Colección Acción Familiar
Ediciones Cinca
N.º 8

Análisis de datos en la investigación social de la familia

fundación acción familiar

Ediciones Cinca

La Suma de Todos

Comunidad de Madrid
www.madrid.org

CONSEJERÍA DE FAMILIA
Y ASUNTOS SOCIALES
Dirección General de Familia

Comunidad de Madrid

Colección Acción Familiar
Ediciones Cinca

Patrocina:

Primera edición: 2008

Diseño cubierta: **Juan Vidaurre**

Producción editorial,
coordinación técnica e impresión:
Grupo editorial CINCA
Avda. Doctor Federico Rubio y Galí, 88
28040 Madrid
Tel. 91 553 22 72. Fax 91 554 37 90
grupoeditorial@edicionescinca.com

Depósito legal: M.
ISBN: 978-84-96889-11-8

Análisis de datos en la investigación social de la familia

Nuria Badenes Plá
M.ª Teresa López López
Carolina Navarro Ruiz
Jorge Onrubia Fernández
César Pérez López
Daniel Santín González
Aurelia Valiño Castro

Dirección y coordinación:

M.ª Teresa López López
Daniel Santín González

fundación acción familiar

Ediciones Cinca

La Suma de Todos

Comunidad de Madrid
www.madrid.org

CONSEJERÍA DE FAMILIA Y ASUNTOS SOCIALES
Dirección General de Familia

Comunidad de Madrid

ÍNDICE

PRÓLOGO

En la actualidad, existe un creciente interés por describir, analizar y comprender todos los aspectos relativos a la realidad social y económica de la familia. Las relaciones entre sus miembros, composición, tamaño y decisiones económicas, son un importante motor para la cohesión económica y social. Su papel es clave para el bienestar social ya que desarrollan funciones imprescindibles en ámbitos tan diversos como: el apoyo en situaciones de dependencia, la prevención de comportamientos no saludables (droga, alcohol, etc), el desarrollo de la solidaridad entre sus miembros, o su importante papel en la lucha contra el fracaso escolar por citar sólo algunos ejemplos.

Poco a poco la mayoría de los gobiernos autonómicos y municipales así como algunas instituciones privadas, están interesadas en ayudar a favorecer todos los aspectos positivos que la familia genera a la sociedad. Y por ello se han ido implementando algunas ayudas tales como: apoyo a la maternidad/paternidad; políticas de rentas mínimas; descuentos a familias numerosas; ayudas económicas en materia educativa; puesta en marcha de centros de atención integral a la familia, entre otras.

La Comunidad de Madrid, consciente de esta necesidad, elaboró su Plan de Apoyo a la Familia que tiene por objetivo articular políticas públicas de familia que respondan fielmente a las necesidades de aquellas que residen en su ámbito territorial, desde la eficiencia y racionalidad económica y social.

Para ayudar a cumplir con este propósito ayudó, con su patrocinio, a la elaboración de este trabajo que ofrece los instrumentos necesarios para facilitar un debate riguroso sobre la familia basado en el análisis de datos, en torno las mejores políticas que en cada momento deben ser puestas en marcha.

Las políticas de familia, lejos de demagogias, requieren un conocimiento analítico previo de la información, con el objetivo de aplicar las mejores medidas de ayuda entre distintas alternativas, aprender de errores y éxitos pasados y comprender la evolución temporal de la realidad de la familia en la sociedad actual. Es en este contexto en el que se presenta este libro que, de manera sencilla y aplicada, expone las principales herramientas matemáticas y estadísticas para extraer información clave sobre la cuál basar la asignación del presupuesto dedicado a la familia.

Esta obra tiene como unidad de análisis la familia pero también a las unidades gestoras del gasto a ella dirigido, y pretende que los conocimientos que se deriven del uso de las distintas técnicas puedan ser fácilmente utilizados por políticos, psicólogos, educadores, sociólogos, economistas, etc.

En definitiva, desde la Fundación Acción Familiar pensamos que la investigación y la toma de decisiones basadas en el rigor científico deberían guiar la puesta en marcha de políticas de ayuda a la familia, huyendo de opiniones sin fundamento analítico, y para ello llevamos trabajando 5 años.

Con este libro esperamos ayudar a todos aquellos que trabajan con y para las familias a promover mejores y más eficientes medidas de apoyo que mejoren su calidad de vida y, por tanto, la de la sociedad.

M.ª Teresa López López
Vicepresidenta
Fundación Acción Familiar

CAPÍTULO I
LA INVESTIGACIÓN SOCIAL DE LAS POLÍTICAS DE FAMILIA

M.ª TERESA LÓPEZ LÓPEZ

1.1. INTRODUCCIÓN

Los principales agentes que participan en la actividad económica son las empresas, el sector público y los hogares, y lo hacen actuando con fuertes interrelaciones entre ellos. Sin embargo es frecuente pensar que las decisiones que toman las familias pertenecen al ámbito estrictamente privado y por tanto no afectan a la sociedad —a su desarrollo económico y estabilidad social—, y mucho menos al sector público —a sus finanzas y equilibrio presupuestario—. Pero en el funcionamiento de una economía de mercado las actuaciones de cada uno de estos agentes están condicionadas y a su vez condicionan, las decisiones tomadas por los otros.

Estas interrelaciones muestran que el Estado no es ajeno a la forma de vida de las familias, y por tanto no debe ser indiferente a las nuevas necesidades a las que éstas se enfrentan. De hecho a los responsables políticos les preocupa, por ejemplo, que las familias decidan tener menos hijos o adopten modos de vida que dificulten o incluso impidan el cuidado de sus mayores. Las finanzas públicas pueden verse alteradas de manera muy importante por estas decisiones que, inicialmente, son concebidas como estrictamente privadas. Que sean privadas no significa que no tengan consecuencias sobre otros individuos e incluso en muchos casos sobre toda la sociedad y sus estructuras económicas. Un país envejecido, en el que no haya un crecimiento suficiente de población, está avocado al fracaso económico y social más absoluto.

Igualmente a las familias no les es indiferente que el sector público les proporcione servicios adecuados, que les permitan tener y cuidar a sus hijos; mantener en sus hogares y bajo sus cuidados a sus mayores; o que permitan a las madres acceder al mercado laboral en igualdad de condiciones con aquellas mujeres que no han tenido hijos. Con toda seguridad estas ayudas serán elementos muy importantes a la hora de tomar las decisiones.

Las empresas tampoco pueden ser ajenas a los nuevos hábitos de consumo y ahorro de las familias que se derivan de sus formas de vida. Éstas se ven obligadas a demandar nuevos servicios que, a su vez, son importantes fuentes de creación de empleo (comida precocinada, servicios de cuidado y atención domiciliaria, etc).

Existen algunos ámbitos en que estas interdependencias son especialmente significativas, relevantes y complejas, tanto a nivel económico como social y que permiten identificar el papel clave que la familia desempeña en el buen funcionamiento de una economía y la convierten en un elemento imprescindible para la cohesión social.

En primer lugar en lo relativo al papel de la familia en el crecimiento económico ya que éste depende fundamentalmente del capital humano y de su formación. La parte más importante de la inversión en éste, y la cobertura de su coste monetario y de oportunidad, se produce en las familias. Esta formación —no sólo académica—, se lleva a cabo fundamentalmente en el seno de la familia, y produce un coste para ésta casi imposible de cuantificar. Los principios de solidaridad y convivencia, la tolerancia, la defensa del hombre, la protección del mas débil, el trabajo en equipo, en definitiva muchos de los valores que humanizan al hombre, se viven y se aprenden, fundamentalmente, en la familia. Una economía que quiere crecer con estabilidad y equilibrio no debe preocuparse solo de tener buenas escuelas y universidades, sino también de ofrecer los medios necesarios para que las personas crezcan en ambientes familiares equilibrados y estables, porque es algo propio de una sociedad sana. En definitiva, y expresándolo en términos estrictamente económicos, en la familia se producen economías externas al permitir una reducción de los costes económicos y sociales, que supone para la sociedad la labor de formación de sus ciudadanos.

En segundo lugar los hogares son unidades de consumo e inversión con pautas de comportamiento muy diferentes a las de los individuos que viven solos. Esto implica que se producen efectos igualmente muy diferentes y que pueden condicionar de manera significativa el comportamiento económico de una sociedad.

En tercer lugar la familia juega un papel clave en la redistribución de la renta y la riqueza. Cuando se tienen hijos, los padres contraen una serie de obligaciones legales que conllevan costes económicos: educación, cuidado y atención de los hijos, etc. De sus resultados y logros se beneficiará toda la sociedad, especialmente en modelos de seguridad social de reparto, ya que las familias con hijos asumen el coste monetario de oportunidad del cuidado, educación y formación de éstos, y serán esos hijos los que en futuro financiarán a través de sus cotizaciones, las pensiones, no solo de sus padres, sino de todos los ciudadanos, hayan tenido o no hijos. Por ello, las familias, junto con el sector público, actúan como unidades que facilitan la redistribución de la renta entre personas y entre generaciones, actuando además como «colchón» de protección en los sistemas de seguridad social.

La familia se configura como una unidad de obligaciones y derechos que favorece el crecimiento económico sostenido, la redistribución de la

renta entre personas y generaciones, estabiliza la sociedad y proporciona una mayor cohesión social, objetivo éste perseguido por el sector público. Todo ello obliga a éste a proporcionar a las familias los instrumentos necesarios para ayudarlas a cumplir sus obligaciones. Sin embargo no vale cualquier política de familia. Las políticas de familia son difíciles de implementar por razones diversas, que se recogen mas adelante, pero sobre todo porque se desarrollan de manera transversal y sus efectos pueden ir mas allá de la simple cobertura de unas necesidades sociales pudiendo llegar a modificar decisiones y pautas de consumo e inversión.

Por todo ello es necesario un análisis riguroso de todas estas interrelaciones y las consecuencias económicas y sociales que pueden producirse, lo que exige avanzar en el conocimiento de todos aquellos instrumentos que nos permitan lograr estos objetivos.

Ayudar a mejorar estos conocimientos es el principal objetivo de las páginas que siguen. Por primera vez, de manera novedosa, se ofrece a los investigadores en temas de familia, un material riguroso sobre un conjunto de herramientas para el análisis de datos, orientados al estudio de la familia y de las políticas que toman a ésta como sujeto beneficiario.

1.2. ALGUNAS CONSIDERACIONES EN TORNO A LA INVESTIGACIÓN EN EL ÁMBITO DE LA FAMILIA

En los últimos años los presupuestos dedicados a servicios sociales y familia han aumentado significativamente. Ello se debe, por un lado, a la mayor demanda social, pero también a una mayor preocupación política por luchar contra las situaciones de necesidad a través de la educación, dependencia, conciliación o la lucha contra la pobreza por citar tan solo algunos ejemplos. Así, en los últimos años han proliferado los Planes de Actuación de los gobiernos y el interés por alcanzar los objetivos que las sociedades del siglo XXI plantean en términos de mayor bienestar y calidad de vida y en los que la familia es la unidad de decisión básica sobre la cual se apoyan estas medidas.

A pesar del aumento presupuestario, las ayudas a la familia se están implantando de forma más lenta de lo que la sociedad demanda aunque todos los grupos políticos de los países desarrollados las incluyan en sus programas. Aunque tradicionalmente estas políticas quedaban en manos del Estado Central, la descentralización administrativa está llevando a que cada vez más regiones y municipios pongan en marcha sus propias iniciativas de ayuda a la familia. Esta lenta pero incesante proliferación de ayudas atomizadas suele tener su base en decisiones de voluntad política mu-

chas veces sin un análisis riguroso previo ni posterior de las consecuencias de su puesta en marcha.

No olvidemos que en un mundo de recursos escasos y de una presión fiscal elevada se debe rendir cuentas a los contribuyentes acerca del uso y los resultados que se está dando a los servicios públicos que éstos financian. En este marco surge la necesidad de cuantificar las necesidades y determinar la población objetivo, elaborar los presupuestos que manejará cada agente gestor, medir resultados de las políticas así como evaluar y tomar decisiones que fortalezcan las intervenciones futuras. Como en cualquier otra ciencia la investigación social debe ser la fuente de alimentación donde acudan políticos y gestores para racionalizar los recursos.

No es el objetivo de esta obra pasar revista a las políticas públicas de familia implementadas en los últimos años [1]. El objetivo de este manual es abordar las principales herramientas y técnicas analíticas para llevar a cabo investigación social y evaluar las políticas sociales y familiares de cualquier organismo público o privado que realice tales tareas.

Las claves de la investigación social de las políticas de familia tienen forma de ciclo que puede ser resumido de la siguiente manera:

a) Medición de las necesidades de partida y caracterización de la población.

b) Establecimiento de los objetivos a alcanzar en función de las necesidades.

c) Establecimiento de los medios y actuaciones o políticas que se utilizarán para su logro dado un presupuesto.

d) Cuantificación de los logros alcanzados por las políticas.

e) Comparación y análisis de los resultados.

f) Establecimiento de las mejores políticas y prácticas y corrección de las menos adecuadas.

Sin embargo, la investigación social y en concreto la evaluación de las políticas públicas de familia presenta algunas peculiaridades propias que dificultan el análisis y a las que deberemos hacer frente.

[1] El lector interesado en conocer las políticas de familia en España y la Unión Europea puede acudir a López *et al.* (2006).

1. Ausencia de precios de referencia. ¿Cuánto *vale* una familia donde los cónyuges son capaces de conciliar vida laboral y familiar?, ¿Cuál es el *precio* para lograr evitar que un adolescente caiga en el consumo de drogas? Muchos de los objetivos y la producción en investigación social no se vende en ningún mercado, lo que dificulta su valoración y cuantificación, aplicándose entonces criterios únicamente subjetivos.

2. Dificultad para identificar a los beneficiarios directos de las medidas. Se trata de actuaciones que, teniendo como destinatarios a los niños, adolescentes o ancianos se implementan a través de sus familias, lo que hace necesario diferenciar entre beneficiarios y usuarios de los servicios o ayudas y esto en ocasiones es complejo.

3. Concurrencia de instrumentos para alcanzar objetivos similares. Las políticas de familia son en muchos casos de carácter transversal con otros servicios sociales y necesitan utilizar instrumentos variados y de carácter complementario.

4. Concurrencia de objetivos a lograr. A diferencia de una empresa privada cuyo objetivo es la maximización del beneficio económico las políticas sociales tienen múltiples objetivos como la ayuda a la infancia, a los adolescentes, a sus familias, conciliación, mediación, etc.

5. Dificultad de identificar y definir previamente indicadores de cantidad y calidad de los beneficiarios y de sus propias familias.

Con estas premisas surge la idea de presentar al analista, gestor e investigador social las principales técnicas estadísticas y econométricas de aplicación en este ámbito. La descripción de las mismas no sólo se realiza a nivel teórico sino que a lo largo del libro se presentan ejercicios resueltos de problemas ilustrativos de la realidad.

1.3. HERRAMIENTAS PARA EL ANÁLISIS DE DATOS

El objetivo de los autores de este manual es poder llegar a que todos los profesionales y lectores interesados en la investigación social sean capaces de replicar en casa los ejemplos y ejercicios que se proponen. Este objetivo añadía dificultad a la elaboración de la obra ya que los paquetes estadísticos y econométricos tradicionales (SPSS, STATA, SAS, MATLAB, EVIEWS, etc.) requieren un esfuerzo inicial en su aprendizaje cuya explicación excedía nuestros objetivos. Es por ello que el software escogido para

desarrollar los ejemplos y aplicaciones es Microsoft ® Excel. Las razones que nos llevaron a escoger Excel como herramienta fueron dos:

1. La elevada difusión de la herramienta Excel en el ámbito de las ciencias sociales que sin duda facilitará el acercamiento del lector al manejo de las técnicas.

2. La sencillez del programa ya que se trabaja en un entorno de hoja de cálculo, sin necesidad de recurrir a comandos de programación. Este hecho permite acercar el análisis de datos a un mayor número de usuarios interesados en las técnicas a nivel aplicado.

Cabe aclarar por tanto que si bien los ejercicios y la discusión aplicada de la obra se implementa utilizando Excel, los contenidos teóricos son útiles para cualquier otro investigador que utilice otro tipo de software. Al lector que desconozca el uso de Excel a un nivel básico-intermedio se le aconseja recibir formación en esta herramienta para un mejor aprovechamiento del libro. Algunas de las técnicas de análisis que se tratan en este manual no podían ejecutarse directamente en la versión estándar de Excel. Es por ello que además de este programa, se utiliza, en algunos de sus capítulos, la herramienta de análisis de datos Addinsoft ® XLSTAT. XLSTAT es un conjunto de utilidades adicionales para Excel que se integran en el programa permitiendo que realice funciones que no vienen programadas por defecto. Este programa puede ser descargado desde la web: http://www.xlstat.com dentro de la pestaña descargar para su evaluación.

Esta obra ha sido redactada para ser utilizada por todo tipo de usuarios (principiantes, intermedios y avanzados). Se pretende ofrecer una descripción intuitiva de lo que hace cada técnica así como de la interpretación de sus resultados pero sin renunciar al desarrollo matemático de cada una de las herramientas utilizadas.

1.4. ESTRUCTURA DEL LIBRO

El libro está estructurado en cuatro bloques cada uno con objetivos bien definidos que a continuación se repasan de forma somera.

BLOQUE I: Selección y Exploración de los Datos

Antes de realizar cualquier investigación social es imprescindible disponer de buenos datos. Ello requiere realizar una recogida de los mismos de forma sistemática utilizando técnicas de muestreo que permitan posteriormente realizar inferencia válida para la población estudiada. Una vez recogidos

los datos resulta también clave realizar una descripción adecuada de todas las variables y gestionar la información perdida y extraña. El éxito de cualquier investigación depende de la correcta obtención, descripción y manipulación inicial de los datos. Este primer bloque comprende los capítulos 2 a 4.

BLOQUE II: Índices de Pobreza, Distribución de la Renta y Eficiencia

En este bloque se trabaja cómo elaborar indicadores que nos permitan evaluar a lo largo del tiempo las actuaciones de las políticas públicas. Dos de los indicadores más usados en muchos ámbitos son los de pobreza y distribución de la renta. De su análisis podremos saber como está siendo la evolución del bienestar social de la población atendida. Dentro de este bloque también se trata el análisis de la eficiencia de las unidades gestoras encargadas en transformar el presupuesto en bienes y servicios al ciudadano. Este bloque comprende los capítulos 5 a 9.

BLOQUE III: Análisis multivariante

El tercer bloque de capítulos abarca las técnicas tradicionales de análisis multivariante aplicadas a la investigación social. Mediante estos análisis es posible construir indicadores complejos, caracterizar los grupos que forman una población o asociar las categorías de distintas variables cualitativas. Mención aparte merece el capítulo 10 dedicado a diseño experimental. En este capítulo se aborda el contraste de hipótesis para determinar si una política pública ha tenido éxito o no sobre la variable cuyo comportamiento queríamos variar. Este bloque comprende los capítulos 10 a 14.

BLOQUE IV: Modelización de datos

En esta última parte se abordan las principales técnicas para el estudio de la causalidad en las ciencias sociales. Por un lado los modelos econométricos de regresión múltiple permiten estudiar la dependencia de una variable continua cuando varía un conjunto de variables explicativas. Por otro lado, el modelo de regresión logística extiende el análisis anterior al caso en el que la variable dependiente es de carácter dicotómico. Este bloque comprende los capítulos 15 y 16.

El contenido de este manual puede ser utilizado en numerosos ámbitos, entre ellos:

- Investigación teórica y aplicada desde un nivel básico (por ejemplo alumnos de licenciatura) a un nivel avanzado (alumnos de maestrías o doctorados).

- Analistas de instituciones públicas y privadas interesadas en la investigación social de todos los aspectos concernientes a la familia.
- Estudiosos de disciplinas como la sociología, economía, psicología, trabajo social o la estadísticas interesados en cuantificar los efectos de las políticas públicas.
- Docencia de métodos de investigación social.

BIBLIOGRAFÍA

López López, M.ª T., Utrilla de la Hoz, A. y Valiño Castro, A. (2006). *Políticas públicas y familia*. Colección Acción Familiar. Ediciones Cinca.

CAPÍTULO II
MUESTREO ESTADÍSTICO

CÉSAR PÉREZ LÓPEZ
DANIEL SANTÍN GONZÁLEZ

2.1. CONCEPTOS INICIALES EN LA TEORÍA DEL MUESTREO

Habitualmente no se dispone de la información censal del conjunto de características familiares e individuales que se pueden estudiar y analizar en una población para investigar cuales son las políticas de familia que más beneficios están aportando o pueden aportar. La razón es que el coste económico de la recogida de toda la información sería muy elevado incluso en municipios pequeños. Es por ello que este problema se resuelve tomando muestras de familias que sean representativas del barrio, municipio, región o país que queremos estudiar. El objetivo de este capítulo es presentar los principales métodos de muestreo en investigación social.

Al hablar de *métodos de muestreo* nos referimos al conjunto de técnicas estadísticas que estudian la forma de seleccionar una *muestra lo suficientemente representativa* de una población cuya información permita inferir las propiedades o características de toda la población cometiendo un *error medible y acotable*. A partir de la muestra, seleccionada mediante un determinado método de muestreo, se estiman las características poblacionales (media, total, proporción, etc.) con un error cuantificable y controlable. Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas *estimadores*, que se convierten en variables aleatorias al considerar la variabilidad de las muestras. Los errores se cuantifican mediante varianzas, desviaciones típicas o errores cuadráticos medios de los estimadores, que miden la precisión de los mismos. La metodología que permite inferir resultados, predicciones y generalizaciones sobre la población estadística, basándose en la información contenida en las muestras representativas previamente elegidas por métodos de muestreo formales, se denomina *inferencia estadística*.

Es muy importante tener en cuenta que para medir el grado de representatividad de la muestra es necesario utilizar *muestreo probabilístico*. Diremos que el muestreo es probabilístico cuando pueda establecerse la probabilidad de obtener cada una de las muestras que sea posible seleccionar, esto es, cuando la selección de muestras constituya un fenómeno aleatorio probabilizable.

Dicha selección se verificará en condiciones de azar, siendo susceptible de medida la incertidumbre derivada de la misma. Esto permitirá medir los errores cometidos en el proceso de muestreo (a través de la varianza u otras medidas estadísticas).

Existen varios tipos de muestreo, dependiendo de que la población estadística sea finita o infinita, materia sobre la que existe amplia literatura estadística, pero nosotros consideraremos solamente el *muestreo en poblaciones finitas* que es el que utilizaremos en la selección de familias e individuos. La población finita inicial que se desea investigar se denomina *población objetivo*, pero el muestreo de toda la población objetivo no siempre es posible debido a diferentes problemas que no permiten obtener información de algunos de sus elementos (inaccesibilidad de algunos de sus elementos, negativas a colaborar, ausencias, etc.), con lo que la población que realmente es objeto de estudio o *población investigada* no suele coincidir con la población objetivo.

Por otro lado, para seleccionar la muestra, necesitaremos un listado de unidades de muestreo denominado *marco* que teóricamente debiera de coincidir con la población objetivo. Un marco será más adecuado cuanto mejor cubra la población objetivo, es decir, cuanto menor sea el *error de cobertura*.

Pero en los marcos son inevitables las desactualizaciones, las omisiones de algunas unidades, las duplicaciones de otras y la presencia de unidades extrañas y otras impurezas que obligan a su depuración (*depuración de marcos imperfectos*). Idealmente podría conseguirse la población objetivo eliminando del marco las unidades erróneamente incluidas en él (unidades extrañas, duplicaciones, etc.) y añadiendo las omisiones. Asimismo, también sería una meta que al eliminar del marco las unidades de las que no se puede obtener información (inaccesibles, ausentes, no colaboradoras, etc.) se obtuviera la población investigada.

El marco puede estar constituido por unidades elementales de muestreo o por unidades compuestas. Una *unidad elemental (o simple)* es la unidad de muestreo más sencilla posible y una *unidad compuesta (o primaria)* está formada por varias unidades elementales. Como en la práctica no es fácil disponer de marcos de unidades elementales, se intentan conseguir marcos de unidades compuestas que son más accesibles.

Por ejemplo, para estudiar habitantes de una región es más fácil disponer de un listado de hogares que de un listado de individuos. Se selecciona la muestra de un marco de hogares (unidades compuestas de varios individuos) y después se estudian las propiedades de los individuos con técnicas adecuadas.

2.2. MUESTREO Y ESTIMADORES. ESTIMACIÓN PUNTUAL

Consideramos la realización de un determinado experimento o fenómeno cuyos resultados se denominan *sucesos*. Entre los experimentos o fenómenos, se denominan *deterministas* aquéllos en los que vamos a conocer *a priori* sus sucesos resultantes, y se denominan *aleatorios* aquéllos cuyos sucesos son desconocidos *a priori*. El estudio de la probabilidad se ocupa de los fenómenos o experimentos aleatorios.

Sean $S_1, S_2, ..., S_n$ los sucesos elementales asociados a un fenómeno o experimento aleatorio dado, entendiendo por *sucesos elementales* los más simples posibles, es decir, aquéllos que no pueden ser descompuestos en otros sucesos. El conjunto $\{S_1, S_2, ..., S_n\}$ se denomina *espacio muestral* asociado al fenómeno o experimento.

Si consideramos como fenómeno o experimento la extracción aleatoria de muestras dentro de una población por un procedimiento o método de muestreo dado, podemos considerar como sucesos elementales las muestras obtenidas, constituyendo el conjunto de las mismas el espacio muestral. Si representamos el conjunto de las N unidades que constituyen la población finita objeto de estudio por $U = \{u_1, u_2, ..., u_N\}$, una muestra de tamaño n puede considerarse como un subconjunto ordenado de n elementos de U, $S_i = \{ u_{i1}, u_{i2}, ..., u_{in}\}$, donde u_{ij} denota el elemento que ocupa el lugar j en en la muestra S_i. Se considera el subconjunto S_i ordenado porque, en general, el orden de colocación de los elementos en las muestras puede ser pertinente, siendo distintas entre sí muestras con los mismos elementos colocados en distinto orden. El conjunto de las N^n muestras posibles de tamaño n que se pueden formar con los N elementos de la población U es el espacio muestral S.

Hay que especificar que, en general, el orden de colocación de los elementos en las muestras sí influye, siendo muestras distintas aquéllas que tienen los mismos elementos situados en distinto orden. Pero lo habitual es que los métodos de muestreo comunes consideren iguales muestras con los mismos elementos, aunque estén colocados en orden diferente. En este caso habitual, en el que el orden de colocación de los elementos en las muestras no se tiene en cuenta, suele expresarse una muestra de tamaño n como $s = \{u_1, u_2, ..., u_n\}$.

Con la finalidad de medir el grado de representatividad de la muestra lo mejor posible, es necesario utilizar muestreo probabilístico. Diremos que *el muestreo es probabilístico* cuando pueda establecerse la probabilidad de obtener cada una de las muestras que sea posible seleccionar (elementos del espacio muestral, S) mediante un procedimiento de muestreo dado, esto es, cuando la selección de muestras constituya un fenómeno ale-

atorio probabilizable. Dicha selección se verificará en condiciones de azar, siendo susceptible de medida la incertidumbre derivada de la misma. Esto permitirá medir los errores cometidos en el proceso de muestreo. Evidentemente, para establecer la probabilidad de todas las muestras posibles derivadas de un procedimiento de muestreo dado, será necesario conocer ese conjunto de muestras, es decir, será necesario delimitar tanto el método de muestreo como el espacio muestral derivado del mismo.

Método de muestreo

Un *procedimiento,* o *método, de muestreo* es sencillamente un proceso o mecanismo mediante el que se seleccionan las muestras de modo que cada una tenga una determinada probabilidad de ser elegida. Por lo tanto, el método aleatorio empleado para seleccionar la muestra define en el espacio muestral *S* una función de probabilidad *P* tal que:

- $P(S_i) \neq 0 \ \forall i$
- $\sum_i P(S_i) = 1$

En general, puede ocurrir que no todas las muestras del espacio muestral pueden ser elegidas. No obstante, consideraremos métodos de muestreo en los que todas las muestras puedan ser seleccionadas, es decir, $P(S_i) > 0 \ \forall i$; se trata entonces de *métodos de muestreo no restringidos*.

En ocasiones suele expresarse un procedimiento de muestreo mediante la terna {*U, S, P*}, que indica que el procedimiento de muestreo definido en la población *U* establece en el espacio muestral *S* asociado la ley de probabilidad *P*.

Método de estimación. Estimadores puntuales

A partir de una muestra, seleccionada mediante un determinado método de muestreo, se estiman las características poblacionales (media, total, proporción, etc.), con un error cuantificable y controlable. Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas *estimadores*, que se convierten en variables aleatorias al considerar la variabilidad de selección de las muestras, y que por tanto cumplen las condiciones de una función de medida. Los errores se cuantifican mediante varianzas, desviaciones típicas o errores cuadráticos medios de los estimadores, que miden la precisión de los mismos.

Supongamos que tenemos definida una característica X en la población $U = \{u_1, u_2, ..., u_N\}$ que toma el valor numérico X_i sobre la unidad u_i $i = 1,2, ..., N$, dando lugar al conjunto de valores $\{X_1, X_2, ... , X_N\}$. Consideramos ahora una cierta función q de los N valores X_i, que suele denominarse parámetro poblacional. Seleccionamos una muestra $s = \{u_1, u_2, ..., u_n\}$ de U mediante un procedimiento de muestreo dado, y consideramos los valores $s(X) = \{X_1, X_2... ,X_n\}$ que toma la característica X en estudio sobre los elementos de la muestra. A partir de estos valores estimamos puntualmente el parámetro poblacional q mediante la expresión $\theta = \theta = (s(X)) = \theta (X_1, ..., X_n)$, basada en los valores X_i $i = 1, 2, ..., n$, que toma la característica X sobre las unidades de la muestra s.

La función θ que asocia a cada muestra s el valor numérico $\theta (s(X)) = \theta (X_1, ..., X_n)$, se denomina *estimador* del parámetro poblacional q. A los valores $\theta (s(X))$ para cada s del espacio muestral se les denomina *estimaciones puntuales*.

Entre los parámetros poblacionales q (función de los N valores X_i) más comunes a estimar, tenemos el total poblacional y la media poblacional para la característica X, definidos de la forma siguiente:

- *Total poblacional:* $X = q (X_1, ..., X_N) = \sum_{i=1}^{N} X_i$
- *Media poblacional:* $\bar{X} = q (X_1, ..., X_N) = \frac{X}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i = \sum_{i=1}^{N} \frac{X_i}{N}$

Hasta ahora hemos supuesto que la característica X definida sobre los elementos de la población es cuantitativa, es decir, cuantificable numéricamente. Sin embargo, también se pueden definir características cualitativas sobre los elementos de la población, como por ejemplo su pertenencia o no a una determinada clase A. Si para cada unidad u_i $i = 1, 2, ..., N$ de la población definimos la característica A_i, que toma valor 1 si la unidad u_i pertenece a la clase A, y que toma valor 0 si la unidad u_i no pertenece a la clase A, podemos definir el total de elementos de la población que pertenecen a la clase A (total de clase) y la proporción de elementos de la población que pertenecen a la clase A (proporción de clase) de la forma siguiente:

- *Total de clase:* $A = q (A_1, ..., A_N) = \sum_{i=1}^{N} A_i$
- *Proporción de clase:* $P = q (A_1, ..., A_N) = \frac{A}{N} = \frac{1}{N}\sum_{i=1}^{N} A_i = \sum_{i=1}^{N} \frac{A_i}{N}$

Analizados ya los cuatro parámetros poblacionales más típicos a estimar, vemos que, en general, un parámetro poblacional q puede expresarse como una suma de elementos Y_i función de los valores que la característica cuantitativa X o cualitativa A considerada toma sobre los elementos de la población. De esta forma, podemos escribir:

$$\theta = \sum_{i=1}^{N} Y_i$$

en cuyo caso tenemos:

$$\begin{cases} Y_i = X_i \text{ para el total poblacional } X \\ Y_i = \dfrac{X_i}{N} \text{ para la media poblacional } \bar{X} \\ Y_i = A_i \text{ para el total de clase } A \\ Y_i = \dfrac{A_i}{N} \text{ para la proporción de clase } P \end{cases}$$

Ahora surge el problema de analizar la forma de los estimadores puntuales óptimos $\theta = \theta (X_1, ..., X_n)$ para estos parámetros poblacionales típicos. Resulta que las mejores propiedades suelen presentarlas los estimadores lineales insesgados de la forma $\hat{\theta} = \sum_{i=1}^{n} w_i Y_i$. Concretamente, *para muestreo sin reposición, el estimador óptimo es el de Horvitz y Thompson* $\hat{\theta}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$, donde p_i es la probabilidad que tiene la unidad u_i de la población de pertenecer a la muestra; *para muestreo con reposición el estimador óptimo es el de Hansen y Hurwitz* $\hat{\theta}_{HH} = \sum_{i=1}^{n} \frac{Y_i}{nP_i}$, donde P_i es la probabilidad de seleccionar la unidad u_i de la población para la muestra (probabilidad unitaria de selección de la unidad u_i).

Existen justificaciones para considerar que el parámetro poblacional $\theta = \sum_{i=1}^{N} Y_i$, puede estimarse convenientemente mediante el estimador $\hat{\theta} = \sum_{i=1}^{n} w_i Y_i$ entre las que podemos citar las siguientes:

- Todas las mediciones de la variable en estudio sobre las unidades de la muestra intervienen en la formación del estimador.

- La importancia de la aportación al estimador de la unidad muestral u_i puede controlarse mediante el coeficiente de ponderación w_i.

- Cuando $w_i = 1$, todas las unidades muestrales intervienen de igual forma en la formación del estimador.

- Los coeficientes w_i pueden depender, entre otros factores, del tamaño de las unidades muestrales, del orden de colocación de las mismas en la muestra, y sobre todo de la probabilidad que tiene la unidad u_i de pertenecer a la muestra según el método de muestreo considerado.

2.3. ESTIMACIÓN POR INTERVALOS

Cuando se realiza una afirmación acerca de los parámetros de la población en estudio basándose en la información contenida en la muestra, bien sea mediante los valores puntuales de un estadístico basado en la misma, bien sea señalando un intervalo de valores dentro del cual se tiene confianza de que esté el valor del parámetro, decimos que estamos ante *estimaciones*. En el primer caso estamos ante el proceso de *estimación puntual,* en el que utilizamos directamente los valores de un estadístico, denominado *estimador puntual,* sobre la muestra dada (*estimaciones puntuales*), para estimar los valores poblacionales. En el segundo caso estamos ante la *estimación por intervalos*, donde se calcula un intervalo de confianza en el que razonablemente cae el valor estimado con un *nivel de confianza* prefijado.

Realizar una estimación por intervalos (o definir un intervalo de confianza) para un parámetro poblacional θ al nivel de confianza α es hallar un intervalo real para el que se tiene una probabilidad $1 - \alpha$ de que el verdadero valor del parámetro θ caiga dentro del citado intervalo. El valor $1 - \alpha$ suele denominarse *coeficiente de confianza*.

Intervalos de confianza cuando el estimador es insesgado

Se trata de estimar el parámetro poblacional mediante un intervalo de confianza basado en el estimador $\hat{\theta}$ insesgado para θ ($E(\hat{\theta})=\theta$). Para estimadores insesgados, es necesario distinguir entre el caso en que la distribución del estimador es normal, y el caso en que dicha distribución no puede asegurarse que sea normal.

a) *El estimador θ tiene una distribución normal*

El intervalo de confianza para el parámetro poblacional q basado en θ será:

$$\left[\hat{\theta}-\lambda_\alpha\sigma\left(\hat{\theta}\right),\hat{\theta}+\lambda_\alpha\sigma\left(\hat{\theta}\right)\right] \text{ con } \lambda_\alpha = F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)$$

F es la función de distribución de la normal (0,1), y a es el nivel de confianza.

Si realmente es dudoso que θ tenga una distribución normal, puede utilizarse la distribución t de Student con $n - 1$ grados de libertad para calcular el intervalo de confianza para q, que en este caso será:

$$\left[\hat{\theta}-t_\alpha\hat{\sigma}\left(\hat{\theta}\right),\hat{\theta}+t_\alpha\hat{\sigma}\left(\hat{\theta}\right)\right] \text{ con } t_\alpha = F^{-1}_{t_{n-1}}\left(1-\frac{\alpha}{2}\right)$$

F es la función de distribución de una t de Student con $n - 1$ grados de libertad.

b) *El estimador θ no tiene una distribución normal*

El intervalo de confianza, derivado de la desigualdad de Tchevichev, para el parámetro poblacional q basado en θ que cubre el valor de q con una probabilidad $1 - a$ (coeficiente de confianza), será:

$$\left[\hat{\theta}-\frac{\sigma\left(\hat{\theta}\right)}{\sqrt{\alpha}},\hat{\theta}+\frac{\sigma\left(\hat{\theta}\right)}{\sqrt{\alpha}}\right]$$

Este intervalo suele ser más ancho que el obtenido cuando la distribución de θ es normal. A medida que θ se aleja más de la normalidad, la anchura de este intervalo es mucho mayor respecto del obtenido para normalidad. Ya sabemos que una estimación por intervalos es tanto mejor cuanto más reducido sea el intervalo de confianza correspondiente, de ahí que la propiedad de normalidad sea muy deseable, pues en este caso los intervalos obtenidos son muy estrechos, lo que implica una buena estimación por intervalos.

Intervalos de confianza en estimadores sesgados

El intervalo de confianza para q basado en el estimador θ en presencia del sesgo no despreciable $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ es el siguiente:

$$\left[\hat{\theta} - \lambda_{\alpha/2}\sigma(\hat{\theta}) - B(\hat{\theta}), \hat{\theta} + \lambda_{\alpha/2}\sigma(\hat{\theta}) - B(\hat{\theta})\right]$$

Observamos que se trata de un intervalo no centrado en θ y desplazado en la cantidad $B(\hat{\theta})$ respecto del intervalo sin sesgo, que debe centrarse situándonos en la peor de las circunstancias, es decir, tomando como extremo fijo del intervalo el más lejano del centro θ, y calculando el otro extremo por equidistancia al centro. Ante esta situación, la presencia del sesgo $B(\hat{\theta})$ origina que el intervalo de confianza para θ basado en el estimador θ y centrado en θ, tenga una longitud superior al intervalo cuando no hay sesgo. Por lo tanto, la presencia de sesgo conduce a una estimación por intervalos menos precisa.

2.4. PRECISIÓN Y COMPARACIÓN DE ESTIMADORES

Como un estimador θ de un parámetro poblacional q es sencillamente una variable aleatoria unidimensional, nos interesarán sus características de centralización y dispersión, particularmente su esperanza, su varianza y sus momentos, así como otras medidas relativas a su precisión.

Precisión de los estimadores

Para analizar la precisión de un estimador suelen utilizarse los conceptos de error de muestreo (o desviación típica), acuracidad (o error cuadrático medio) y sesgo. Suele llamarse precisión a la acuracidad, lo que no es del todo correcto, ya que, aunque la acuracidad sea la magnitud más general para la medición de la precisión, hay casos en los que el análisis puede realizarse en función de otras magnitudes, como el sesgo o la desviación típica. Todas estas magnitudes que influyen en la precisión de un estimador pueden relacionarse a partir de la *descomposición del error cuadrático medio en sus componentes* de la forma siguiente:

$$ECM(\hat{\theta}) = \sigma(\hat{\theta})^2 + B(\hat{\theta})^2$$

Por tanto, la acuracidad (error cuadrático medio) de un estimador se descompone en la suma del cuadrado del error de muestreo y el cuadrado del sesgo.

En la práctica, se considera que el sesgo de θ no es influyente cuando:

$$\left|\frac{B(\hat{\theta})}{\sigma(\hat{\theta})}\right| < \frac{1}{10}$$

Comparación de estimadores

Comparación de estimadores insesgados

Un estimador θ insesgado para el parámetro poblacional θ tiene la propiedad de que su error cuadrático medio coincide con su varianza, ya que al ser $E(\hat{\theta}) = \theta$ se tiene:

$$V(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2 = E(\hat{\theta} - \theta)^2 = ECM(\hat{\theta})$$

De esta forma los conceptos de acuracidad y error del estimador son similares para estimadores insesgados. Por tanto, *para comparar varios estimadores insesgados* θ_i *del parámetro poblacional* θ en cuanto a precisión

bastará considerar sus errores de muestreo $\sigma(\hat{\theta}_i) = +\sqrt{V(\hat{\theta}_i)}$ siendo más preciso el estimador que menor error de muestreo presente.

También en el caso de insesgadez el concepto de error relativo de muestreo puede expresarse en términos de una única magnitud variable $\sigma(\hat{\theta})$ ya que:

$$CV(\hat{\theta}) = \frac{\sigma(\hat{\theta})}{E(\hat{\theta})} = \frac{\sigma(\hat{\theta})}{\theta}$$

y al ser q una constante el error relativo está en función sólo del error de muestreo.

Con lo que resulta que, en el caso de estimadores insesgados, la precisión puede hacerse depender exclusivamente del error de muestreo $\sigma(\hat{\theta})$.

Comparación de estimadores sesgados

Para estimadores θ sesgados del parámetro poblacional θ, la magnitud general para analizar su precisión es su error cuadrático medio. Por tanto, para comparar varios estimadores sesgados del parámetro poblacional θ en cuanto a precisión se utilizará el error cuadrático medio y el estimador más preciso será el que menor error cuadrático medio presente.

Pero en la práctica el cálculo del error cuadrático medio puede ser problemático. Por esta razón *cuando se intentan comparar varios estimadores* $\hat{\theta}_i$ *del parámetro poblacional* q *todos sesgados*, se calcula para cada uno de ellos la cantidad:

$$\left|\frac{B(\hat{\theta}_i)}{\sigma(\hat{\theta}_i)}\right|$$

siendo más preciso aquel estimador que presenta una relación del sesgo al error de muestreo en valor absoluto más pequeña. También puede utilizarse el coeficiente de variación $CV(\hat{\theta}_i)=\sigma(\hat{\theta}_i)/E(\hat{\theta}_i)$, siendo más preciso el estimador con menor coeficiente de variación (error relativo). Se observa que el denominador del coeficiente de variación es el valor esperado del estimador, con lo que el coeficiente de variación recoge el efecto de un posible sesgo en el estimador.

Si los estimadores sesgados a comparar tienen todos sesgo despreciable, es decir $|B(\hat{\theta}_i)/\sigma(\hat{\theta}_i)|<1/10$, se compararían como si fuesen insesgados, de acuerdo a lo expresado en el apartado anterior.

Comparación de estimadores sesgados e insesgados

Para comparar en cuanto a precisión varios estimadores $\hat{\theta}_i$ unos sesgados y otros insesgados del parámetro poblacional θ, se utilizará el error cuadrático medio, y el estimador más preciso será el que menor error cuadrático medio presente. A veces, ante las dificultades de cálculo del error

cuadrático medio se utiliza el coeficiente de variación $CV(\hat{\theta}_i) = \sigma(\hat{\theta}_i)/E(\hat{\theta}_i)$ (que contempla el posible efecto del sesgo en su denominador), siendo más preciso el estimador con menor coeficiente de variación (error relativo).

Si los estimadores sesgados tienen todos sesgo despreciable $|B(\hat{\theta}_i)/\sigma(\hat{\theta}_i)| < 1/10$, se haría la comparación global como insesgados de acuerdo a los valores de $\sigma(\hat{\theta}_i)$.

Cuantificación de la ganancia en precisión de los estimadores

Para medir la precisión de los estimadores suele utilizarse el error cuadrático medio, el error relativo (coeficiente de variación) o el error de muestreo (desviación típica). En cada caso, la ganancia en precisión vendrá dada por las respectivas tasas de variación:

$$\left(\frac{ECM(\hat{\theta}_1)}{ECM(\hat{\theta}_2)} - 1\right) x100 \quad \left(\frac{CV(\hat{\theta}_1)}{CV(\hat{\theta}_2)} - 1\right) x100 \quad \left(\frac{\sigma(\hat{\theta}_1)}{\sigma(\hat{\theta}_2)} - 1\right) x100$$

2.5. MUESTREO ALEATORIO SIMPLE SIN REPOSICIÓN. ESTIMADORES Y ERRORES

Se trata de un procedimiento de selección de muestras con probabilidades iguales, que consiste en obtener la muestra unidad a unidad de forma aleatoria sin reposición a la población de las unidades previamente seleccionadas, teniendo presente además que el orden de colocación de los elementos en las muestras no interviene (es decir, que muestras con los mismos elementos colocados en orden distinto se consideran iguales).

De esta forma, las muestras con elementos repetidos son imposibles. Como el procedimiento de selección es con probabilidades iguales, todas las muestras son equiprobables, y además se cumple que todas las unidades de la población tienen la misma probabilidad de pertenecer a la muestra $p_i = n/N$.

Supongamos en todo momento que el tamaño de la población es N y el tamaño de la muestra es n. Como la muestra se selecciona sin reposi-

ción, se realiza la selección sucesiva de las unidades para la muestra con probabilidades *1/(N - t)* para valores de *t = 0, 1,..., n*.

Los estimadores para las características poblacionales más interesantes son los siguientes:

$$\theta = X = \sum_{i=1}^{N} X_i \Rightarrow Y_i = X_i \Rightarrow \hat{\theta} = \hat{X} = \sum_{i=1}^{n} \frac{X_i}{\pi_i} = \sum_{i=1}^{n} \frac{X_i}{\frac{n}{N}} = N \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i}_{\bar{x}} = N\bar{x}$$

$$\theta = \bar{X} = \sum_{i=1}^{N} \frac{X_i}{N} \Rightarrow Y_i = \frac{X_i}{N} \Rightarrow \hat{\theta} = \hat{\bar{X}} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{\pi_i} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{x}$$

$$\theta = P = \sum_{i=1}^{N} \frac{A_i}{N} \Rightarrow Y_i = \frac{A_i}{N} \Rightarrow \hat{\theta} = \hat{P} = \sum_{i=1}^{n} \frac{\frac{A_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} A_i = \hat{P}$$

$$\theta = A = \sum_{i=1}^{N} A_i \Rightarrow Y_i = A_i \Rightarrow \hat{\theta} = \hat{A} = \sum_{i=1}^{n} \frac{A_i}{\frac{n}{N}} = N \frac{1}{n} \sum_{i=1}^{n} A_i = N\hat{P}$$

Se observa que el estimador lineal insesgado de la media poblacional es la media muestral, el estimador del total poblacional es el tamaño poblacional por la media muestral, el estimador de la proporción poblacional es la proporción muestral, y el estimador del total de clase poblacional es el tamaño poblacional por la proporción muestral.

Los errores de estos estimadores, en términos de sus varianzas, son los siguientes:

$$V_{ms}\left(\hat{\bar{X}}\right)=(1-f)\frac{S^2}{n}=(1-\frac{n}{N})\frac{\frac{N}{N-1}\sigma^2}{n}=\frac{N-n}{N-1}\frac{\sigma^2}{n}$$

$$V_{ms}\left(\hat{X}\right)=N^2(1-f)\frac{S^2}{n}=N^2\frac{N-n}{N-1}\frac{\sigma^2}{n}$$

$$V_{ms}\left(\hat{A}\right)=N^2(1-\frac{n}{N})\frac{\frac{N}{N-1}PQ}{n}=N^2\frac{N-n}{N-1}\frac{PQ}{n}$$

$$V_{ms}\left(\hat{P}\right)=(1-\frac{n}{N})\frac{\frac{N}{N-1}PQ}{n}=\frac{N-n}{N-1}\frac{PQ}{n}$$

Las estimaciones de estos errores son las siguientes:

$$\hat{V}_{ms}\left(\hat{\bar{X}}\right)=(1-f)\frac{\hat{S}^2}{n}=(1-\frac{n}{N})\frac{\frac{n}{n-1}\hat{\sigma}^2}{n}=\frac{N-n}{N}\frac{\hat{\sigma}^2}{n-1}$$

$$\hat{V}_{ms}\left(\hat{X}\right)=N^2(1-f)\frac{\hat{S}^2}{n}=N(N-n)\frac{\hat{\sigma}^2}{n-1}$$

$$\hat{V}_{ms}\left(\hat{A}\right)=N(N-n)\frac{\hat{P}\hat{Q}}{n-1}$$

$$V_{ms}\left(\hat{P}\right)=\frac{N-n}{N}\frac{\hat{P}\hat{Q}}{n-1}$$

2.6. MUESTREO ALEATORIO SIMPLE CON REPOSICIÓN. ESTIMADORES Y ERRORES

Se trata de un procedimiento de selección con probabilidades iguales, que consiste en obtener la muestra unidad a unidad de forma aleatoria con reposición a la población de las unidades previamente seleccionadas. De esta forma las muestras con elementos repetidos son posibles, y cualquier elemento de la población puede estar repetido en la muestra *0, 1, ..., n* veces.

Supongamos en todo momento que el tamaño de la población es N y el tamaño de la muestra es n. Como la muestra se selecciona con reposición (se reponen a la población las unidades previamente seleccionadas) y con probabilidades iguales, se realiza la selección sucesiva de las unidades para la muestra con probabilidades $P_i = 1/N$ y todas las muestras son equiprobables, ya que:

$$P(u_1, u_2, \ldots, u_n) = P(u_1)P(u_2) \ldots P(u_n) = (1/N)(1/N) \ldots (1/N) = 1/(N^n)$$

Se observa que se obtienen los mismos estimadores insesgados para los parámetros poblacionales que en el caso de muestreo aleatorio simple sin reposición, tal y como se indica a continuación:

$$\theta = X = \sum_{i=1}^{N} X_i \Rightarrow Y_i = X_i \Rightarrow \hat{\theta} = \hat{X} = \sum_{i=1}^{n} \frac{X_i}{nP_i} = \sum_{i=1}^{n} \frac{X_i}{\frac{n}{N}} = N \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i}_{\bar{x}} = N\bar{x}$$

$$\theta = \bar{X} = \sum_{i=1}^{N} \frac{X_i}{N} \Rightarrow Y_i = \frac{X_i}{N} \Rightarrow \hat{\theta} = \hat{\bar{X}} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{nP_i} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{x}$$

$$\theta = P = \sum_{i=1}^{N} \frac{A_i}{N} \Rightarrow Y_i = \frac{A_i}{N} \Rightarrow \hat{\theta} = \hat{P} = \sum_{i=1}^{n} \frac{\frac{A_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} A_i = \hat{P}$$

$$\theta = A = \sum_{i=1}^{N} A_i \Rightarrow Y_i = A_i \Rightarrow \hat{\theta} = \hat{A} = \sum_{i=1}^{n} \frac{A_i}{\frac{n}{N}} = N \frac{1}{n} \sum_{i=1}^{n} A_i = N\hat{P}$$

Se observa que los estimadores coinciden con los del caso sin reposición. Los errores de estos estimadores (en términos de sus varianzas) y sus estimaciones son los siguientes:

$$V_{ca}\left(\hat{\bar{X}}\right) = \frac{\sigma^2}{n} \qquad \hat{V}_{ca}\left(\hat{\bar{X}}\right) = \frac{\hat{S}^2}{n}$$

$$V_{ca}\left(\hat{\bar{X}}\right) = N^2 \frac{\sigma^2}{n} \qquad \hat{V}_{ca}\left(\hat{\bar{X}}\right) = N^2 \frac{\hat{S}^2}{n}$$

$$V_{ca}\left(\hat{A}\right) = N^2 \frac{PQ}{n} \qquad \hat{V}_{ca}\left(\hat{A}\right) = N^2 \frac{\hat{P}\hat{Q}}{n-1}$$

$$V_{ca}\left(\hat{P}\right) = \frac{PQ}{n} \qquad \hat{V}_{ca}\left(\hat{P}\right) = \frac{\hat{P}\hat{Q}}{n-1}$$

2.7. TAMAÑO DE LA MUESTRA EN M.A.S.

Estudiaremos el *tamaño de muestra necesario para cometer un error de muestreo* (absoluto, relativo o con coeficiente de confianza adicional) dependiendo de si se estima la media, el total, la proporción o el total de clase y de si hay o no reposición.

Muestreo aleatorio simple sin reposición

En el caso del *tamaño de muestra necesario para cometer un error absoluto de muestreo* $e = s(\theta\ )$ dependiendo de si θ estima la media, el total, la proporción o el total de clase, se despeja n de la expresión $e = s(\theta\ )$ para un e dado. Tenemos:

Media:

$$e = \sigma(\hat{\bar{X}}) = \sqrt{(1-f)\frac{S^2}{n}} \Rightarrow e^2 = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = \frac{S^2}{n} - \frac{S^2}{N}$$

$$\Rightarrow \frac{S^2}{n} = e^2 + \frac{S^2}{N} \Rightarrow n = \frac{S^2}{e^2 + \frac{S^2}{N}} = \frac{NS^2}{Ne^2 + S^2}$$

Total:

$$e = \sigma(\hat{X}) = \sqrt{N^2(1-f)\frac{S^2}{n}} \Rightarrow e^2 = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n} = \frac{N^2S^2}{n} - \frac{N^2S^2}{N} \Rightarrow$$

$$\Rightarrow \frac{N^2S^2}{n} = e^2 + \frac{N^2S^2}{N} \Rightarrow n = \frac{N^2S^2}{e^2 + \frac{N^2S^2}{N}} = \frac{N^2S^2}{\frac{Ne^2 + N^2S^2}{N}} = \frac{N^2S^2}{e^2 + NS^2}$$

Proporción:

En las expresiones anteriores el valor de S^2 es desconocido y debe ser aproximado fundamentalmente a partir de dos estrategias. La primera consiste en utilizar el valor calculado en otros trabajos que midan el mismo concepto y otro parecido. Mediante la segunda estrategia se tomaría una muestra pequeña de forma previa al muestreo total para realizar el cálculo de S^2 y tener así una idea aproximada de su valor. Si sustituimos el valor de S^2 para variables A_i (que sólo toman los valores 0 y 1) en la fórmula del tamaño muestral para la media tendremos para la estimación de la proporción el tamaño:

$$n = \frac{NS^2}{Ne^2 + S^2} = \frac{N\frac{N}{N-1}PQ}{\frac{N}{N-1}PQ + Ne^2} = \frac{N^2PQ}{\frac{NPQ + (N-1)Ne^2}{N}} = \frac{NPQ}{e^2(N-1) + PQ}$$

En el caso de la proporción se observa que cuando $N \to \infty$ (fracción de muestreo n/N tendiendo a cero) el tamaño muestral $n \to S^2/e^2$. Pero:

$$S^2/e^2 = \frac{N}{N-1}PQ\Big/e^2 \cong PQ/e^2 = n_0$$

Es decir, n inversamente proporcional al cuadrado del error de muestreo y directamente proporcional a la proporción poblacional P. En este caso, la misma precisión da una muestra de tamaño n para una población de N elementos que para una población de N' elementos con $N' > N$ siempre y cuando se cumpla la desigualdad definida por:

$$N > n_0(n_0 - 1) = \frac{\frac{N}{N-1}PQ}{e^2}\left(\frac{\frac{N}{N-1}PQ}{e^2} - 1\right) \cong \frac{PQ}{e^2}\left(\frac{PQ}{e^2} - 1\right)$$

Para la estimación de la proporción es muy interesante tener en cuenta que para poblaciones grandes o fracción de muestreo pequeña ($N \to \infty$), el valor máximo de n se obtiene para $P = Q = 1/2$. Para constatar este resultado sabemos que si $N \to \infty$ el tamaño muestral n tiende al valor $n_0 = PQ/e^2 = f(P)$, expresión que tenemos que maximizar en P. Si igualamos la primera derivada al valor cero tenemos que como $f(P) = P(1-P)/e^2$ entonces $f'(P) = (1-2P)/e^2 = 0 \quad P = 1/2$. Por otra parte $f''(P) = -2/e^2 < 0$, lo que asegura la presencia de un máximo para la función f en el punto $P = 1/2$. Como $Q = 1-P = 1-1/2 = 1/2$, el valor máximo de n para poblaciones grandes o fracciones de muestreo pequeñas se obtiene para $P = Q = 1/2$. Por lo tanto, para un error prefijado se necesitarán tamaños de muestra más pequeños cuanto más próximo esté P a cero o a uno. Este resultado es muy importante en la práctica, ya que *cuando se estiman proporciones y no se conoce el valor de la proporción poblacional P ni se tiene una aproximación suya (proporcionada por una encuesta similar, por una encuesta piloto, por la misma encuesta realizada anteriormente o por cualquier otro método), entonces se toma P=1/2*, con lo que estamos situándonos en el caso de máximo tamaño muestral para el error fijado, lo cual siempre es aceptable estadísticamente. La dificultad práctica puede ser que se obtenga un tamaño muestral n demasiado grande para el presupuesto de que se dispone.

Total de clase:

Si sustituimos el valor de S^2 para variables A_i (que sólo toman los valores 0 y 1) en la fórmula del tamaño muestral para el total tendremos el tamaño:

$$n=\frac{N^2S^2}{e^2+NS^2}=\frac{N^2\frac{N}{N-1}PQ}{e^2+\frac{N}{N-1}PQN}=\frac{N^2PQ}{e^2(N-1)+N^2PQ}$$

También puede estudiarse el *tamaño de muestra necesario para cometer un error relativo de muestreo* $e_r = Cv(\theta)$ dependiendo de si se estima la media, el total, la proporción y el total de clase. Asimismo, es típico introducir un coeficiente de confianza adicional P_a al error de muestreo a cometer (*límite de tolerancia*). En este caso las fórmulas de los *tamaños muestrales necesarios para cometer un error absoluto o relativo de muestreo dado en presencia del coeficiente de confianza adicional* se derivarán de las expresiones $e_a = \lambda_a s(\theta)$ y $e_{ra} = \lambda_a Cv(\theta)$. En general $\lambda_a = F^{-1}(1-a/2)$, siendo F la función de distribución de una normal (0,1). El cuadro siguiente resume las expresiones de los tamaños muestrales.

Tipo de error → Parámetro ↓	Absoluto e	Relativo e_r	Absoluto y coeficiente de confianza adicional e_a	Relativo y confianza e_{ra}
Media	$\frac{NS^2}{Ne^2+S^2}$	$\frac{NC_{1,a}^2}{Ne_r^2+C_{1,a}^2}$	$\frac{\lambda_a^2NS^2}{Ne^2+\lambda_a^2S^2}$	$\frac{\lambda_a^2NC_{1,a}^2}{Ne_{ra}^2+\lambda_a^2C_{1,a}^2}$
Total	$\frac{N^2S^2}{e^2+NS^2}$	$\frac{NC_{1,a}^2}{Ne_r^2+C_{1,a}^2}$	$\frac{\lambda_a^2N^2S^2}{e^2+\lambda_a^2NS^2}$	$\frac{\lambda_a^2NC_{1,a}^2}{Ne_{ra}^2+\lambda_a^2C_{1,a}^2}$
Proporción	$\frac{NPQ}{e^2(N-1)+PQ}$	$\frac{NQ}{P(N-1)e_r^2+Q}$	$\frac{\lambda_a^2NPQ}{e^2(N-1)+\lambda_a^2PQ}$	$\frac{NQ\lambda_a^2}{e_{ra}^2(N-1)P+\lambda_a^2Q}$
Total de clase	$\frac{N^2PQ}{e^2(N-1)+N^2PQ}$	$\frac{NQ}{P(N-1)e_r^2+Q}$	$\frac{\lambda_a^2N^2PQ}{e^2(N-1)+\lambda_a^2N^2PQ}$	$\frac{NQ\lambda_a^2}{e_{ra}^2(N-1)P+\lambda_a^2Q}$

En todas las fórmulas S^2 es la cuasivarianza poblacional y $C_{1,a}^2=\left(S/\overline{X}\right)^2$. Por otra parte, es el valor crítico de la normal unitaria al nivel a.

Muestreo aleatorio simple con reposición

Igual que en el caso de sin reposición, consideraremos el *tamaño de muestra necesario para cometer un error de muestreo* $e = s(\theta)$ dependien-

do de si θ estima la media, el total, la proporción o el total de clase. También se considerará el *tamaño de muestra necesario para cometer un error relativo de muestreo* $e_r = Cv(\theta)$ dependiendo de si se estima la media, el total, la proporción y el total de clase. Asimismo, se tendrá presente la introducción de un coeficiente de confianza adicional P_a al error de muestreo a cometer (*límite de tolerancia*), en cuyo caso las fórmulas de los *tamaños muestrales necesarios para cometer un error absoluto o relativo de muestreo dado en presencia del coeficiente de confianza adicional* se derivarán de las expresiones $e_a = \lambda_a s(\theta)$ y $e_{ra} = \lambda_a Cv(\theta)$. En general, $\lambda_a = F^{-1}(1-a/2)$, siendo F la función de distribución de una normal (0,1). El cuadro siguiente resume las expresiones de los tamaños muestrales.

Tipo de error → Parámetro ↓	Absoluto e	Relativo e_r	Absoluto y coeficiente de confianza adicional e_{aa}	Relativo y coeficiente e_{ra}
Media	$\frac{\sigma^2}{e^2}$	$\frac{C_x^2}{e_r^2}$	$\frac{\lambda_a^2\sigma^2}{e^2}$	$\frac{\lambda_a^2 C_x^2}{e_{ra}^2}$
Total	$\frac{N^2\sigma^2}{e^2}$	$\frac{C_x^2}{e_r^2}$	$\frac{\lambda_a^2 N^2\sigma^2}{e^2}$	$\frac{\lambda_a^2 C_x^2}{e_{ra}^2}$
Proporción	$\frac{PQ}{e^2}$	$\frac{Q}{Pe_r^2}$	$\frac{\lambda_a^2 PQ}{e^2}$	$\frac{\lambda_a^2 Q}{Pe_{ra}^2}$
Total de clase	$\frac{N^2PQ}{e^2}$	$\frac{Q}{Pe_r^2}$	$\frac{\lambda_a^2 N^2PQ}{e^2}$	$\frac{\lambda_a^2 Q}{Pe_{ra}^2}$

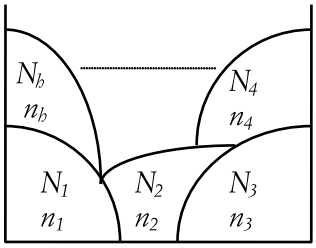
En todas las fórmulas s^2 es la varianza poblacional y $C_x^2 = \left(\sigma/\bar{X}\right)^2$. Por otra parte, es el valor crítico de la normal unitaria al nivel a.

2.8. MUESTREO ESTRATIFICADO. ESTIMADORES Y ERRORES

En el muestreo estratificado, una *población heterogénea* con N unidades $\{u_i\}_{i=1,2,\ldots,N}$ se subdivide en L *subpoblaciones lo más homogéneas posibles* no solapadas, denominadas *estratos* $\{u_{hi}\}_{\substack{h=1,2,\ldots,L \\ i=1,2,\ldots,N_h}}$ de tamaños $N_1, N_2, \ldots, N_L$.

La muestra estratificada de tamaño n se obtiene seleccionando n_h elementos ($h = 1, 2, \ldots, L$) de cada uno de los L estratos en los que se subdivide la población de forma independiente. Si la muestra estratificada se obtiene seleccionando una muestra aleatoria simple en cada estrato de forma independiente, el muestreo se *denomina muestreo aleatorio estratificado*, pero en general nada impide utilizar diferentes tipos de selección en cada estrato. Si el muestreo aleatorio en cada estrato es sin reposición el muestreo estratificado es sin reposición, y si el muestreo aleatorio en cada estrato es con reposición el muestreo estratificado es con reposición.

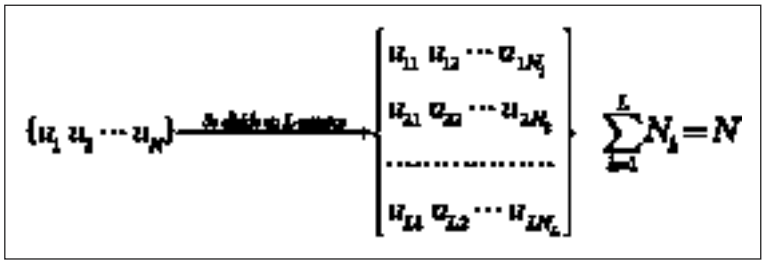
Podemos representar gráficamente la población dividida en h estratos de tamaño N_h; de cada uno de ellos seleccionamos de modo independiente n_h unidades (mediante muestreo aleatorio simple si no se especifica otra cosa) para la muestra estratificada de tamaño n, de la forma siguiente:



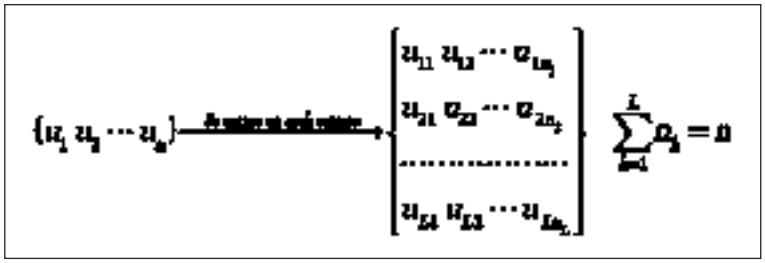
POBLACIÓN

Podemos expresar la formación de estratos en la población y la formación de la muestra estratificada de la forma siguiente:

POBLACIÓN



MUESTRA



Estimadores y errores

Los estimadores lineales insesgados de los parámetros poblacionales clásicos en muestreo estratificado, tanto con reposición como sin reposición, son los siguientes

$$\hat{X}_e = \sum_{h=1}^{L} N_h \hat{\bar{X}}_h = \sum_{h=1}^{L} N_h \bar{x}_h = \sum_{h=1}^{L} \hat{X}_h \quad \hat{\bar{X}}_e = \bar{x}_e = \sum_{h=1}^{L} \underbrace{\frac{N_h}{N}}_{W_h} \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^{L} W_h \bar{x}_h$$

$$\hat{A}_e = \sum_{h=1}^{L} N_h \hat{P}_h = \sum_{h=1}^{L} \hat{A}_h \quad \hat{P}_e = \sum_{h=1}^{L} W_h \hat{P}_h$$

El estimador del total poblacional en muestreo estratificado aleatorio es la suma de los estimadores del total en cada estrato. El estimador de la media en muestreo estratificado aleatorio es la media ponderada de los estimadores de la media en cada estrato, siendo los coeficientes de ponderación $W_h = N_h/N$ de suma unitaria. El estimador del total de clase en muestreo estratificado aleatorio es la suma de los estimadores del total de clase en cada estrato. El estimador de la proporción en muestreo estratificado aleatorio es la media ponderada de los estimadores de la proporción en cada estrato, siendo los coeficientes de ponderación $W_h = N_h/N$ de suma unitaria.

Las varianzas de los estimadores y sus errores son los siguientes:

$$V(\hat{X}_e) = V\left(\sum_{h=1}^{L} \hat{X}_h\right) = \sum_{h=1}^{L} V(\hat{X}_h) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

$$V(\bar{x}_e) = V\left(\sum_{h=1}^{L} W_h \bar{x}_h\right) = \sum_{h=1}^{L} W_h^2 V(\bar{x}_h) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

$$V(\hat{A}_e) = V\left(\sum_{h=1}^{L} \hat{A}_h\right) = \sum_{h=1}^{L} V(\hat{A}_h) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{N_h}{N_h - 1} \frac{P_h Q_h}{n_h}$$

$$V(\hat{P}_e) = V\left(\sum_{h=1}^{L} W_h \hat{P}_h\right) = \sum_{h=1}^{L} W_h^2 V(\hat{P}_h) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{N_h}{N_h - 1} \frac{P_h Q_h}{n_h}$$

$$\hat{V}(\hat{X}_e) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\bar{X}_e) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{A}_e) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{n_h}{n_h - 1} \frac{\hat{P}_h \hat{Q}_h}{n_h} = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{\hat{P}_h \hat{Q}_h}{n_h - 1}$$

$$\hat{V}(\hat{P}_e) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{\hat{P}_h \hat{Q}_h}{n_h - 1}$$

Para el caso del muestreo estratificado con reposición los estimadores son los mismos, y sus varianzas son las siguientes:

$$V(\hat{X}_e)=V\left(\sum_{h=1}^{L}\hat{X}_h\right)=\sum_{h=1}^{L}V(\hat{X}_h)=\sum_{h=1}^{L}N_h^2\frac{\sigma_h^2}{n_h}$$

$$V(\bar{x}_e)=V\left(\sum_{h=1}^{L}W_h\bar{x}_h\right)=\sum_{h=1}^{L}W_h^2V(\bar{x}_h)=\sum_{h=1}^{L}W_h^2\frac{\sigma_h^2}{n_h}$$

$$V(\hat{A}_e)=V\left(\sum_{h=1}^{L}\hat{A}_h\right)=\sum_{h=1}^{L}V(\hat{A}_h)=\sum_{h=1}^{L}N_h^2\frac{P_hQ_h}{n_h}$$

$$V(\hat{P}_e)=V\left(\sum_{h=1}^{L}W_h\hat{P}_h\right)=\sum_{h=1}^{L}W_h^2V(\hat{P}_h)=\sum_{h=1}^{L}W_h^2\frac{P_hQ_h}{n_h}$$

Las estimaciones de los errores (estimaciones de varianzas) son las siguientes:

$$\hat{V}(\hat{X}_e)=\sum_{h=1}^{L}N_h^2\frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\bar{X}_e)=\sum_{h=1}^{L}W_h^2\frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{A}_e)=\sum_{h=1}^{L}N_h^2\frac{n_h}{n_h-1}\frac{\hat{P}_h\hat{Q}_h}{n_h}=\sum_{h=1}^{L}N_h^2\frac{\hat{P}_h\hat{Q}_h}{n_h-1}$$

$$\hat{V}(\hat{P}_e)=\sum_{h=1}^{L}W_h^2\frac{\hat{P}_h\hat{Q}_h}{n_h-1}$$

2.9. AFIJACIÓN DE LA MUESTRA. VARIANZAS Y SU ESTIMACIÓN SIN REPOSICIÓN

Se llama afijación de la muestra al reparto, asignación, adjudicación, adscripción o distribución del tamaño muestral n entre los diferentes estratos; esto es, a la determinación de los valores de n_h que verifiquen $n_1 + n_2 + ... + n_L = n$. Pueden establecerse muchas afijaciones o maneras de repartir la muestra entre los estratos, pero las más importantes son: la afijación uniforme, la afijación proporcional, la afijación de varianza mínima y la afijación óptima.

Afijación uniforme

Consiste en asignar el mismo número de unidades muestrales a cada estrato, con lo que se tomarán todos los n_h iguales a n/L, aumentando o

disminuyendo este tamaño en una unidad si n no fuese múltiplo de L, esto es, $n_h = E(n/L) + 1$, donde E denota la parte entera.

$$n_h = k \;\; \forall h = 1 \cdots L \Rightarrow \sum_{h=1}^{L} n_h = \sum_{h=1}^{L} k \Rightarrow n = Lk \Rightarrow f_h = \frac{n_h}{N_h} = \frac{k}{N_h}$$

Para este tipo de afijación, las varianzas de los estimadores y sus estimaciones se hallan sustituyendo en las fórmulas generales f_h por k/N_h. Este tipo de afijación da la misma importancia a todos los estratos, en cuanto a tamaño de la muestra, con lo cual favorecerá a los estratos de menor tamaño y perjudicará a los grandes en cuanto a precisión. Sólo es conveniente en poblaciones con estratos de tamaño similar.

Afijación proporcional

Consiste en asignar a cada estrato un número de unidades muestrales proporcional a su tamaño. Las n unidades de la muestra se distribuyen proporcionalmente a los tamaños de los estratos expresados en número de unidades. Tenemos:

$$n_h = N_h k \Rightarrow \underbrace{\sum_{h=1}^{L} n_h}_{n} = \sum_{h=1}^{L} N_h k = k \underbrace{\sum_{h=1}^{L} N_h}_{N} \Rightarrow n = kN \Rightarrow k = \frac{n}{N} = f$$

$$f_h = \underbrace{\frac{n_h}{N_h}}_{n_h} = \frac{N_h k}{N_h} = k = f \quad W_h = \frac{N_h}{N} = \frac{n_h/k}{n/k} = \frac{n_h}{n}$$

Para este tipo de afijación, las varianzas de los estimadores serán:

$$V(\bar{X}_{st}) = \frac{(1-k)}{k} \sum_{h=1}^{L} N_h \cdot S_h^2 \;\; , \;\; V(\bar{x}_{st}) = \frac{(1-k)}{n} \sum_{h=1}^{L} W_h \cdot S_h^2$$

$$V(\hat{A}_{st}) = \frac{(1-k)}{k} \sum_{h=1}^{L} \frac{N_h^2}{N_h - 1} \cdot P_h Q_h, \;\; V(\hat{P}_{st}) = \frac{(1-k)}{k} \sum_{h=1}^{L} \frac{N_h^2/N}{N_h - 1} \cdot P_h Q_h$$

En afijación proporcional los estimadores de media y total pueden expresarse como sigue:

$$\hat{X}_{st} = \sum_{h=1}^{L} N_h \bar{x}_h = \sum_{h=1}^{L} \frac{n_h}{k} \bar{x}_h = \frac{1}{K} \sum_{h=1}^{L} n_h \underbrace{\bar{x}_h}_{x_h / n_h} = \frac{\sum_{h=1}^{L} x_h}{k} = \frac{x}{f} = \frac{Total\ muestral}{Fracción\ de\ muestreo}$$

$$\hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = \sum_{h=1}^{L} \frac{n_h}{n} \bar{x}_h = \frac{1}{n} \sum_{h=1}^{L} n_h \underbrace{\bar{x}_h}_{x_h / n_h} = \frac{\sum_{h=1}^{L} x_h}{n} = \frac{Total\ muestral}{Tamaño\ de\ muestra}$$

A la vista de los resultados anteriores, en afijación proporcional, podemos asegurar lo siguiente:

- Las fracciones de muestreo en los estratos son iguales y coinciden con la fracción global de muestreo, siendo su valor la constante de proporcionalidad.

- Los coeficientes de ponderación W_h se obtienen exclusivamente a partir de la muestra, pues para su cálculo sólo son necesarios valores muestrales (n_h y n).

- El estimador insesgado para el total poblacional puede expresarse como el cociente entre el total muestral y la fracción de muestreo, o lo que es lo mismo, como el producto del total muestral por la inversa de la fracción de muestreo. Similar propiedad tiene el estimador insesgado para el total de clase (producto del total de clase muestral por la inversa de la fracción de muestreo).

- El estimador insesgado para la media poblacional puede expresarse como el cociente entre el total muestral y el tamaño de la muestra. Similar propiedad tiene el estimador insesgado para la proporción poblacional (cociente entre el total de clase muestral y el tamaño de la muestra).

- Como $\pi_{hi} = \frac{n_h}{N_h} = \frac{k}{K} = f$ todas las unidades de la población tienen la misma probabilidad de figurar en la muestra de n unidades; es decir, estamos en el caso de muestras autoponderadas.

Afijación de mínima varianza (o afijación de Neyman)

La afijación de mínima varianza o afijación de Neyman consiste en determinar los valores de n_h (número de unidades que se extraen del estrato

h-ésimo para la muestra) de forma que para un tamaño de muestra fijo igual a *n* la varianza de los estimadores sea mínima.

La expresión para n_h es $n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} = n \cdot \frac{\frac{N_h}{N} S_h}{\sum_{h=1}^{L} \frac{N_h}{N} S_h} = n \cdot \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$

Vemos que los valores de n_h son proporcionales a los productos y en el supuesto de que $S_h = S$, $\forall h = 1, 2, ..., L$ esta afijación de mínima varianza coincidiría con la proporcional, tal y como se ve a continuación:

$$S_h = S \Rightarrow n_h = n \cdot \frac{N_h S}{\sum_{h=1}^{L} N_h S} = \frac{n N_h}{N} = k N_h \text{ con } k = \frac{n}{N}$$

La utilidad de esta afijación es mayor si hay grandes diferencias en la variabilidad de los estratos. En otro caso, la mayor sencillez y autoponderación de la afijación proporcional hacen preferible el empleo de ésta.

Una vez calculados los n_h para afijación de mínima varianza, vamos a ver cuánto vale la *varianza del estimador de la media y del total* para este tipo de afijación. Tenemos:

$$V(\bar{x}_*) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2, \quad V(\hat{X}_*) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h\right)^2 - \sum_{h=1}^{L} N_h S_h^2$$

Si se quiere la afijación y la expresión de la *varianza mínima para el estimador de la proporción y el total de clase,* basta sustituir en la fórmula anterior S_h^2 por $P_h Q_h N_h/(N_h\text{-}1)$.

Afijación óptima

La afijación óptima consiste en determinar los valores de n_h (número de unidades que se extraen del estrato *h*-ésimo para la muestra) de forma que para un coste fijo *C* la varianza de los estimadores sea mínima. El coste fijo *C* será la suma de los costes derivados de la selección de las unidades muestrales de los estratos; es decir, si c_h es el coste por unidad de muestreo en el estrato *h*, el coste total de selección de las n_h unidades muestrales en ese estrato será $c_h n_h$. Sumando los costes $c_h n_h$ para los *L* estratos tenemos el coste total de selección de la muestra estratificada.

Podemos escribir que $$n_h = n \cdot \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}} = n \cdot \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}$$

Vemos que los valores de n_h son proporcionales a los productos $N_h \cdot S_h / \sqrt{c_h}$ y en el supuesto de que $C_h = k \ \forall h = 1, 2, ..., L$ (coste constante en todos los estratos) la afijación óptima coincide con la de mínima varianza, y si además $S_h = S$, $\forall h = 1, 2, ..., L$ la afijación óptima coincidirá con la de mínima varianza y con la proporcional.

Valor de la varianza mínima

Una vez calculados los n_h para afijación óptima, vamos a ver cuánto vale la *varianza del estimador de la media y del total* para este tipo de afijación. Tenemos:

$$V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

$$V(\hat{X}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} N_h S_h \sqrt{c_h}\right) - \sum_{h=1}^{L} N_h S_h^2$$

Si se quiere la afijación óptima y la expresión de la *varianza mínima para el estimador de la proporción y el total de clase,* basta sustituir en la fórmula anterior S_h^2 por $P_h Q_h N_h/(N_h\text{-}1)$.

2.10. AFIJACIÓN DE LA MUESTRA. VARIANZAS Y SU ESTIMACIÓN CON REPOSICIÓN

Dada la forma en que están definidos los cálculos de los n_h para las afijaciones uniforme y proporcional, dichas afijaciones no van a verse afectadas por el hecho de que el muestreo sea con o sin reposición. Sin embargo, sí variarán las varianzas de los estimadores. Las afijaciones de mínima varianza y óptima sí van a verse afectadas por la existencia de reposición o no, ya que el cálculo de n_h depende de las varianzas en los estratos.

Afijación uniforme

Para este tipo de afijación, las varianzas de los estimadores serán:

$$V(\hat{X}_{st})=\sum_{h=1}^{L}N_h^2\frac{\sigma_h^2}{k}, \quad V(\bar{x}_{st})=\sum_{h=1}^{L}W_h^2\frac{\sigma_h^2}{k},$$
$$V(\hat{A}_{st})=\sum_{h=1}^{L}N_h^2\frac{P_hQ_h}{k}, \quad V(\hat{P}_{st})=\sum_{h=1}^{L}W_h^2\frac{P_hQ_h}{k}$$

Afijación proporcional

Para este tipo de afijación las varianzas de los estimadores serán:

$$V(\hat{X}_{st})=\frac{1}{k}\sum_{h=1}^{L}N_h\sigma_h^2, \quad V(\hat{A}_{st})=\frac{1}{k}\sum_{h=1}^{L}N_hP_hQ_h,$$
$$V(\bar{x}_{st})=\frac{1}{n}\sum_{h=1}^{L}W_h\sigma_h^2, \quad V(\hat{P}_{st})=\frac{1}{n}\sum_{h=1}^{L}W_h\frac{P_hQ_h}{k}$$

Afijación de mínima varianza (o afijación de Neyman)

Tenemos:

$$n_h=n\cdot\frac{W_h\sigma_h}{\sum_{h=1}^{L}W_h\sigma_h}=n\cdot\frac{N_h\sigma_h}{\sum_{h=1}^{L}N_h\sigma_h}, \quad V(\bar{x}_{st})=\frac{1}{n}\left(\sum_{h=1}^{L}W_h\sigma_h\right)^2,$$
$$V(\bar{x}_{st})=\frac{1}{n}\left(\sum_{h=1}^{L}N_h\sigma_h\right)^2$$

Si se quiere la *afijación de mínima varianza y la expresión de la varianza mínima para el estimador de la proporción y el total de clase* basta sustituir en la fórmula anterior s_h^2 por P_hQ_h.

Afijación óptima

Tenemos:

$$n_h=n\cdot\frac{\frac{W_h\sigma_h}{\sqrt{c_h}}}{\sum_{h=1}^{L}\frac{W_h\sigma_h}{\sqrt{c_h}}}=n\cdot\frac{\frac{N_h\sigma_h}{\sqrt{c_h}}}{\sum_{h=1}^{L}\frac{N_h\sigma_h}{\sqrt{c_h}}}$$

$$V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h \sigma_h \sqrt{c_h}\right)$$

$$V(X_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h \sigma_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} N_h \sigma_h \sqrt{c_h}\right)$$

Si se quiere la *afijación óptima y la expresión de la varianza mínima para el estimador de la proporción y el total de clase* basta sustituir en las fórmulas anterioriores s_h^2 por $P_h Q_h$.

2.11. TAMAÑO DE LA MUESTRA

Muestreo estratificado sin reposición

Vamos a analizar ahora el tamaño de muestra estratificada necesario para cometer un determinado error de muestreo conocido de antemano. Distinguiremos los casos de error de muestreo dado con y sin coeficiente de confianza adicional y, además, distinguiremos entre los diferentes tipos de afijación de la muestra.

Tipo de error → Parámetro ↓	Absoluto proporcional	Absoluto varianza mínima	Absoluto y coeficiente de confianza adicional proporcional	Absoluto y coeficiente de confianza adicional varianza mínima
Media	$\dfrac{\sum_{h=1}^{L} W_h S_h^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$	$\dfrac{\sum_{h=1}^{L} W_h S_h^2}{\frac{e^2}{\lambda_p^2} + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{\frac{e^2}{\lambda_p^2} + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$
Total	$\dfrac{N\sum_{h=1}^{L} N_h S_h^2}{e^2 + \sum_{h=1}^{L} N_h S_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{e^2 + \sum_{h=1}^{L} N_h S_h^2}$	$\dfrac{N\sum_{h=1}^{L} N_h S_h^2}{\frac{e^2}{\lambda_p^2} + \sum_{h=1}^{L} N_h S_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{\frac{e^2}{\lambda_p^2} + \sum_{h=1}^{L} N_h S_h^2}$
Proporción	$\dfrac{\sum_{h=1}^{L} W_h \frac{N_h}{N_h - 1} P_h Q_h}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h \frac{N_h}{N_h - 1} P_h Q_h}$	$\dfrac{\left(\sum_{h=1}^{L} W_h \sqrt{\frac{N_h}{N_h - 1} P_h Q_h}\right)^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h \frac{N_h}{N_h - 1} P_h Q_h}$	$\dfrac{\sum_{h=1}^{L} W_h \frac{N_h}{N_h - 1} P_h Q_h}{\frac{e^2}{\lambda_p^2} + \frac{1}{N}\sum_{h=1}^{L} W_h \frac{N_h}{N_h - 1} P_h Q_h}$	$\dfrac{\left(\sum_{h=1}^{L} W_h \sqrt{\frac{N_h}{N_h - 1} P_h Q_h}\right)^2}{\frac{e^2}{\lambda_p^2} + \frac{1}{N}\sum_{h=1}^{L} W_h \frac{N_h}{N_h - 1} P_h Q_h}$
Total de clase	$\dfrac{N\sum_{h=1}^{L} N_h \frac{N_h}{N_h - 1} P_h Q_h}{e^2 + \sum_{h=1}^{L} N_h \frac{N_h}{N_h - 1} P_h Q_h}$	$\dfrac{\left(\sum_{h=1}^{L} N_h \sqrt{\frac{N_h}{N_h - 1} P_h Q_h}\right)^2}{e^2 + \sum_{h=1}^{L} N_h \frac{N_h}{N_h - 1} P_h Q_h}$	$\dfrac{N\sum_{h=1}^{L} N_h \frac{N_h}{N_h - 1} P_h Q_h}{\frac{e^2}{\lambda_p^2} + \sum_{h=1}^{L} N_h \frac{N_h}{N_h - 1} P_h Q_h}$	$\dfrac{\left(\sum_{h=1}^{L} N_h \sqrt{\frac{N_h}{N_h - 1} P_h Q_h}\right)^2}{\frac{e^2}{\lambda_p^2} + \sum_{h=1}^{L} N_h \frac{N_h}{N_h - 1} P_h Q_h}$

Muestreo estratificado con reposición

Vamos a analizar ahora el tamaño de muestra estratificada con reposición necesario para cometer un determinado error de muestreo conocido de antemano. Distinguiremos los casos de error de muestreo dado con y sin coeficiente de confianza adicional y, además, distinguiremos entre los diferentes tipos de afijación de la muestra.

Tipo de error Parámetro	*Absoluto proporcional*	*Absoluto varianza máxima*	*Absoluto y coeficiente de confianza adicional proporcional*	*Absoluto y coeficiente de confianza adicional varianza máxima*
Media	$\frac{\sum_{h=1}^{L} W_h S_h^2}{e^2}$	$\frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{e^2}$	$\frac{\sum_{h=1}^{L} W_h S_h^2}{e^2/z^2}$	$\frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{e^2/z^2}$
Total	$\frac{N \sum_{h=1}^{L} N_h S_h^2}{e^2}$	$\frac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{e^2}$	$\frac{N \sum_{h=1}^{L} N_h S_h^2}{e^2/z^2}$	$\frac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{e^2/z^2}$
Proporción	$\frac{\sum_{h=1}^{L} W_h P_h Q_h}{e^2}$	$\frac{\left(\sum_{h=1}^{L} W_h \sqrt{P_h Q_h}\right)^2}{e^2}$	$\frac{\sum_{h=1}^{L} W_h P_h Q_h}{e^2/z^2}$	$\frac{\left(\sum_{h=1}^{L} W_h \sqrt{P_h Q_h}\right)^2}{e^2/z^2}$
Total de clase	$\frac{N \sum_{h=1}^{L} N_h P_h Q_h}{e^2}$	$\frac{\left(\sum_{h=1}^{L} N_h \sqrt{P_h Q_h}\right)^2}{e^2}$	$\frac{N \sum_{h=1}^{L} N_h P_h Q_h}{e^2/z^2}$	$\frac{\left(\sum_{h=1}^{L} N_h \sqrt{P_h Q_h}\right)^2}{e^2/z^2}$

2.12. MUESTREO SISTEMÁTICO. ESTIMADORES Y ERRORES

Consideramos una población de tamaño *N,* y agrupamos sus elementos en *n* zonas de tamaño *k* ($N = nk$). Para extraer una muestra de tamaño *n* se elige al azar una unidad en la primera zona, y para seleccionar las $n - 1$ unidades restantes para la muestra se toma en cada zona la unidad que ocupa el mismo lugar dentro de su zona que el que ocupaba la primera unidad seleccionada dentro de la primera zona. Por ejemplo, si la unidad seleccionada para la muestra al azar en la primera zona es la tercera, se elegirán las $n - 1$ unidades restantes para la muestra tomando la tercera unidad de cada zona. Las muestras sistemáticas así obtenidas suelen denominarse *muestras 1 en k.*

Se demuestra que un estimador lineal insesgado para la media poblacional es la media de la muestra sistemática obtenida, para la proporción poblacional es la proporción de la muestra sistemática, para el total poblacional es *N* veces el total de la muestra sistemática, y para el total de

clase es N veces el total de clase muestral. Es decir, podemos escribir lo siguiente:

- *Total* $\hat{X} = N\bar{x}_i$
- *Media* $\hat{\bar{X}} = \bar{x}_i$
- *Proporción* $\hat{P} = \hat{P}_i$
- *Total de clase* $\hat{A} = N\hat{P}_i$

A partir de la tabla del análisis de la varianza para la población que se presenta a continuación, pueden calcularse los errores de los estimadores.

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre muestras	k 1	$\sum_{i}^{n}\sum_{j}^{k}(\bar{x}_i \ \bar{X})^2$	S_b^2
Dentro de muestras	N k	$\sum_{i}^{n}\sum_{j}^{k}(X_{ij} \ \bar{x}_i)^2$	S_w^2
Total	k 1+(N k)=N 1	$\sum_{i}^{n}\sum_{j}^{k}(X_{ij} \ \bar{X}_i)$	S^2

$$V(\hat{\bar{X}}) = V(\bar{x}_i) = (1 \ f)\frac{S_w^2}{n}, \quad V(\hat{X}) = V(N\bar{x}_i) = N^2V(\bar{x}_i) = N^2(1 \ f)\frac{S_w^2}{n}$$

$$V(\hat{P}) = V(\hat{P}_i) = \frac{1}{k}\sum_{j}^{k}(\hat{P}_j \ P)^2 = \frac{1}{nk}\sum_{i}^{n}\sum_{j}^{k}(\hat{P}_j \ P)^2 = \frac{1}{N}\sum_{i}^{n}\sum_{j}^{k}(\hat{P}_j \ P)^2$$

$$V(\hat{A}) = V(N\hat{P}_i) = N^2V(\hat{P}_i) = N^2\frac{1}{k}\sum_{j}^{k}(\hat{P}_j \ P)^2 = N\sum_{i}^{n}\sum_{j}^{k}(\hat{P}_j \ P)^2$$

Un concepto interesante en muestreo sistemático es el coeficiente de correlación intramuestral r_w, que mide la interrelación entre las unidades dentro de las muestras. Lógicamente, esta interrelación debe de ser lo más pequeña posible, ya que en el muestreo sistemático interesa la heterogeneidad intramuestral, con la finalidad de que una única muestra sistemática represente lo mejor posible a toda la población. Para que una muestra sistemática aspire a ser fiel espejo de toda la población ha de ser heterogénea, y la interrelación entre sus unidades ha de ser baja. Por lo tanto, inicialmente parece

lógico que interesen valores muy pequeños del coeficiente de correlación intramuestral. La expresión matemática de r_w es la siguiente:

$$= \frac{2 \quad (X_i \ \bar{X})(X_i \ \bar{X})}{N\Omega \ 1)^{\ 2}}, \text{con} \quad ^{2} = \frac{1}{nk} \quad (X_i \ \bar{X})^{2} = \text{varianza poblacional}$$

La varianza de los estimadores puede expresarse en función de r_w. Para la media tenemos:

$$V(\bar{x}_i) = \frac{N \ n}{N} \frac{S^2}{n} \ 1+(n \ 1)$$

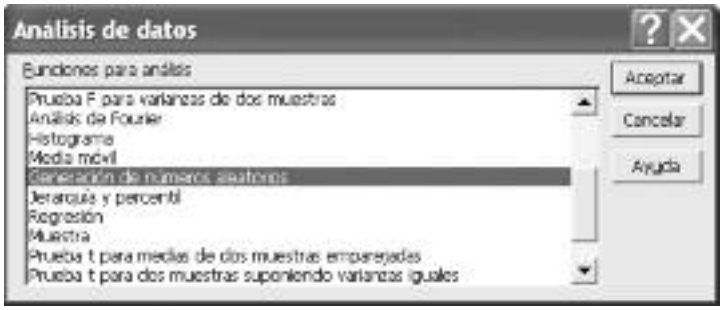


FIGURA 2-1

FIGURA 2-2

Según esta expresión, la precisión del muestreo sistemático puede analizarse en función del coeficiente de correlación intramuestral, de tal modo que la precisión máxima se produce para $r_w = -1/(n+1)$, y la mínima para $r_w = 0$, *igualándose la precisión del muestreo sistemático con la del muestreo aleatorio simple para* $r_w = 0$. De esta forma, para valores de r_w entre $-1/(n+1)$ y 0, el muestreo sistemático es más preciso que el aleatorio simple, y para valores de r_w entre 0 y 1, el muestreo sistemático es menos preciso que el aleatorio simple. Por lo tanto, en cuanto a precisión, convienen valores negativos del coeficiente de correlación intraconglomerados r_w.

La relación anterior permite estimar las varianzas en el muestreo sistemático mediante las fórmulas del aleatorio simple cundo r_w se aproxima a cero.

2.13. FUNCIONES Y HERRAMIENTAS DE EXCEL PARA MUESTRAS ALEATORIAS

Excel dispone de dos funciones para selección de números aleatorios uniformemente con reposición. Su sintaxis es la siguiente:

`ALEATORIO ( )`	*Devuelve un número aleatorio mayor o igual que 0 y menor que 1, distribuido uniformemente. Cada vez que se calcula la hoja de cálculo, se devuelve un número aleatorio nuevo. Si desea usar ALEATORIO para generar un número aleatorio, pero no desea que los números cambien cada vez que se calcule la celda, puede escribir =ALEATORIO ( ) en la barra de fórmulas, y después presionar la tecla F9 para cambiar la fórmula a un número aleatorio. Para generar un número real aleatorio entre a y b, use ALEATORIO ( )*(b-a)+a.*
`ALEATORIO.ENTRE (a,b)`	*Devuelve un número aleatorio entre los números a y b.*

Por otra parte, Excel permite obtener números aleatorios independientes, extraídos según una distribución dada, utilizando herramientas de análisis. Si en el cuadro de diálogo *Análisis de datos* de la Figura 2-1 elegimos *Generación de números aleatorios*, se obtiene el cuadro de diálogo *Generación de números aleatorios* de la Figura 2-2. En el cuadro *Números de variables* introduzca el número de columnas de valores que desee incluir en la tabla de resultados. Si no introduce ningún número, Microsoft Excel rellenará todas las columnas del rango de salida que se haya especificado. En el cuadro *Cantidad de números aleatorios* introduzca el número de puntos de datos que desee ver. Cada punto de datos aparecerá en una fila de la tabla de resultados. Si no introduce ningún número, Microsoft Excel re-

llenará todas las columnas del rango de salida que se haya especificado. En el cuadro *Distribución* haga clic en la distribución estadística que desee utilizar para crear los valores aleatorios.

Las distribuciones posibles son:

Uniforme: Caracterizada por los límites inferior y superior. Se extraen las variables con probabilidades iguales de todos los valores del rango. Una aplicación normal utilizará una distribución uniforme en el rango 0...1.

	A
1	-2,5043119
2	0,34869686
3	1,3207341
4	0,81364988
5	-2,3642497
6	-0,3806656
7	-2,6107955
8	0,04671847
9	0,03416289
10	0,13331032

FIGURA 2-3

Normal: Caracterizada por una media y una desviación estándar. Una aplicación normal utilizará una media de 0 y una desviación estándar de 1 para la distribución estándar normal.

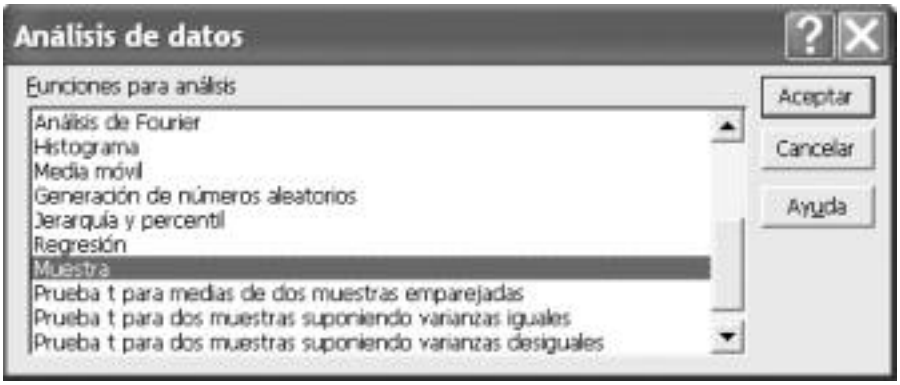


FIGURA 2-4

FIGURA 2-5

	A	B	C	D
1	-2,5043119	1	9	
2	0,34869686	14	11	
3	1,3207341	35	9	
4	0,81364988	57	57	
5	-2,3642497	68	19	
6	-0,3808656	9	15	
7	-2,6107955	12	68	
8	0,04671847	13	17	
9	0,03416289	15	10	
10	0,13331032	6	12	
11		2	1	
12		3	6	
13		4	57	
14		5	12	
15		8	9	
16		10		
17		11		
18		15		
19		16		
20		17		
21		18		
22		19		

FIGURA 2-6

Bernoulli: Caracterizada por la probabilidad de éxito (valor *p*) en un ensayo dado. Las variables aleatorias de Bernoulli tienen el valor 0 ó 1; por ejemplo, puede trazarse una variable aleatoria uniforme en el rango 0...1. Si la variable es menor o igual que la probabilidad de éxito, se asignará el valor 1 a la variable aleatoria de Bernoulli; en caso contrario, se le asignará el valor 0.

Binomial: Caracterizada por una probabilidad de éxito (valor *p*) durante un número de pruebas; por ejemplo, se pueden generar variables aleatorias de Bernoulli de número de pruebas, cuya suma será una variable aleatoria binomial.

Poisson: Caracterizada por un valor *lambda*, igual a *1/media*. La distribución de Poisson se utiliza con frecuencia para caracterizar el número de incidencias por unidad de tiempo; por ejemplo, el ritmo promedio al que llegan los vehículos a una garita de peaje.

Frecuencia relativa: Caracterizada por un límite inferior y superior, un incremento, un porcentaje de repetición para valores y un ritmo de repetición de la secuencia.

Discreta: Caracterizada por un valor y el rango de probabilidades asociado. El rango debe contener dos columnas. La columna izquierda deberá contener valores, y la derecha probabilidades asociadas con el valor de esa fila. La suma de las probabilidades deberá ser 1.

En el campo *Parámetros* introduzca un valor o varios valores para caracterizar la distribución seleccionada. En el campo *Iniciar con* escriba un valor opcional a partir del cual se generarán números aleatorios. Podrá volver a utilizar este valor para generar los mismos números aleatorios más adelante. En el cuadro *Rango de salida* introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados; Microsoft Excel determinará el tamaño del área de resultados, y mostrará un mensaje si la tabla de resultados reemplaza datos ya existentes. Haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para asignar un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado. En la Figura 2-3, se muestra la salida correspondiente a las opciones de *Generación de números aleatorios* de la Figura 2-2 (10 números aleatorios normales de media cero y varianza 1 con semilla 50).

Adicionalmente, Excel permite obtener una muestra aleatoria simple con reposición de una población numérica dada como rango de entrada. Si en el

cuadro de diálogo *Análisis de datos* de la Figura 2-4 elegimos *Muestra*, se obtiene el cuadro de diálogo *Muestra* de la Figura 2-5. A continuación, se explica la funcionalidad de todos los campos del cuadro de diálogo *Muestra*.

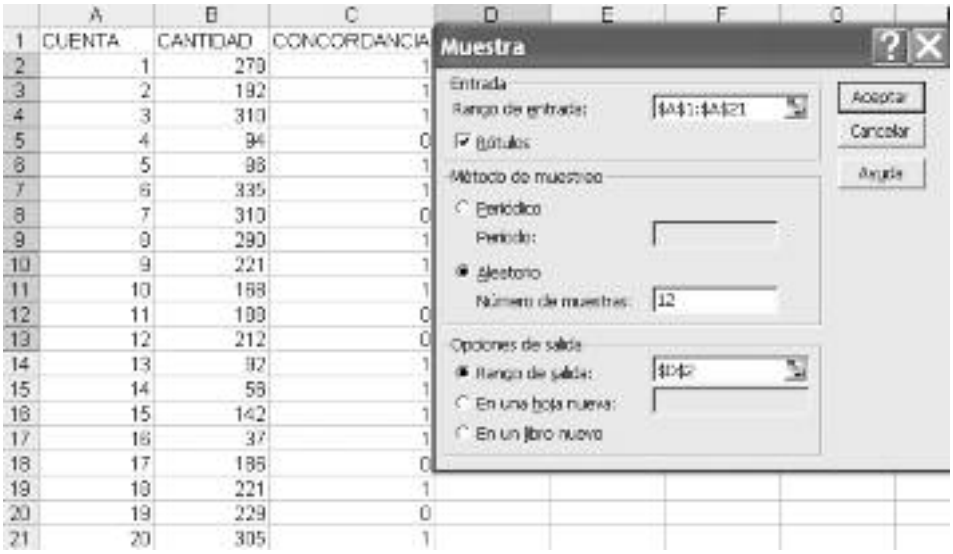
FIGURA 2-7

EJERCICIO 2-1. Un encuestador muestrea aleatoriamente 20 encuestas a familias con dos hijos, verificando en 12 de ellas su consumo quincenal en euros, así como si los dos progenitores trabajan (1) o no (0). Se tiene la siguiente estructura poblacional:

Cuenta	Cantidad	Concordancia	Cuenta	Cantidad	Concordancia
1	278	1	11	188	0
2	192	1	12	212	0
3	310	1	13	92	1
4	94	0	14	56	1
5	86	1	15	142	1
6	335	1	16	37	1
7	310	0	17	186	0
8	290	1	18	221	1
9	221	1	19	229	0
10	168	1	20	305	1

Basándose en las 12 familias verificadas, estimar la proporción de familias cuyos dos miembros trabajan, así como el importe medio de consumo mensual, y cuantificar el error cometido.

Rango de entrada: Introduzca la referencia correspondiente al rango de datos que contenga la población de valores de los que desee extraer una mues-



	A	B	C
1	CUENTA	CANTIDAD	CONCORDANCIA
2	1	279	1
3	2	192	1
4	3	310	1
5	4	94	0
6	5	98	1
7	6	335	1
8	7	310	0
9	8	290	1
10	9	221	1
11	10	168	1
12	11	109	0
13	12	212	0
14	13	92	1
15	14	58	1
16	15	142	1
17	16	37	1
18	17	185	0
19	18	221	1
20	19	229	0
21	20	305	1

FIGURA 2-8

tra. Microsoft Excel extraerá muestras de la primera columna, luego de la segunda, y así sucesivamente.

Rótulos: Active esta casilla si la primera fila y la primera columna del rango de entrada contienen rótulos. Desactívela si el rango de entrada carece

	A	B	C	D	E	F
1	CUENTA	CANTIDAD	CONCORDANCIA	MUESTRA	X	A
2	1	279	1	17	185	0
3	2	192	1	10	221	1
4	3	310	1	4	94	0
5	4	94	0	5	98	1
6	5	98	1	1	279	1
7	6	335	1	6	335	1
8	7	310	0	2	192	1
9	8	290	1	7	310	0
10	9	221	1	14	92	1
11	10	168	1	20	305	1
12	11	109	0	3	310	1
13	12	212	0	16	142	1
14	13	92	1	Estimaciones:	=PROMEDIO(X)	=PROMEDIO(A)
15	14	58	1	Errores absolutos:	=VARP(CANTIDAD)/12	=VARP(CONCORDANCIA)/12
16	15	142	1	Errores relativos:	=100*RAIZ(E15)/E14	=100*RAIZ(F15)/F14
17	16	37	1			
18	17	185	0			
19	18	221	1			
20	19	229	0			
21	20	305	1			

FIGURA 2-9

	A	B	C	D	E	F
1	CUENTA	CANTIDAD	CONCORDANCIA	MUESTRA	X	A
2	1	278	1	17	186	0
3	2	192	1	18	221	1
4	3	310	1	4	94	0
5	4	94	0	5	86	1
6	5	86	1	1	278	1
7	6	335	1	6	335	1
8	7	310	0	2	192	1
9	8	290	1	7	310	0
10	9	221	1	14	56	1
11	10	188	1	20	305	1
12	11	188	0	3	310	1
13	12	212	0	15	142	1
14	13	92	1	**Estimaciones:**	**209,5833333**	**0,75**
15	14	56	1	**Errores absoluto**	**655,745**	**0,0175**
16	15	142	1	**Errores relativos**	**12,21829905**	**17,6383421**
17	16	37	1			
18	17	186	0			
19	18	221	1			
20	19	229	0			
21	20	305	1			

FIGURA 2-10

de rótulos; Excel generará los rótulos de datos correspondientes para la tabla de resultados.

Método de muestreo: Haga clic en *Periódico* o *Aleatorio* para indicar el intervalo de muestreo que desee.

EJERCICIO 2-2. El Director de un Centro de Atención a la Familia desea estimar el tiempo promedio que necesita un trabajador para atender una llamada telefónica. El CAF tiene 98 operadores, y se selecciona una muestra de 8, a los que se les toma el tiempo, obteniéndose los siguientes resultados:

4,2 5,1 7,9 3,8 5,3 4,6 5,1 4,1

Estimar el tiempo promedio y el tiempo total para realizar la tarea entre todos los operadores, estableciendo límites al 95% para los errores de estimación.

Período: Introduzca el intervalo periódico en el que desee realizar la muestra. El valor *n* del período del rango de entrada y cada valor *n* del perío-

do siguiente se copiarán en la columna de resultados. El muestreo terminará cuando se llegue al final del rango de entrada.

Número de muestras: Introduzca el número de valores aleatorios que desee en la columna de resultados. Cada valor se extrae de una posición aleatoria del rango de entrada, y puede seleccionarse cualquier número más de una vez.

Rango de salida: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Los datos se escribirán en una sola columna debajo de la celda. Si selecciona *Periódico*, el número de valores de la tabla de resultados será igual al número de valores del rango de entrada dividido por la tasa de muestreo. Si selecciona *Aleatorio*,

FIGURA 2-11

	A	B	C	D
1	T		T	
2	4,2			
3	5,1		Media	5,0125
4	7,9		Error típico	0,45451603
5	3,8		Mediana	4,65
6	5,3		Moda	5,1
7	4,6		Desviación estándar	1,28556547
8	5,1		Varianza de la muestra	1,65287857
9	4,1		Curtosis	4,24373466
10	MEDIA ESTIMADA=		Coeficiente de asimetría	1,87830493
11	TOTAL ESTIMADO=		Rango	4,1
12			Mínimo	3,8
13			Máximo	7,9
14			Suma	40,1
15			Cuenta	8
16			Nivel de confianza(95,0%)	1,07475008
17				

FIGURA 2-12

	A	B	C	D
1	T		TI	
2	4,2			
3	5,1		Media	5,0125
4	7,9		Error típico	0,454516029891783
5	3,8		Mediana	4,85
6	5,3		Moda	5,1
7	4,6		Desviación estándar	1,28558546757731
8	5,1		Varianza de la muestra	1,65287057142058
9	4,1		Curtosis	4,24373466005059
10	MEDIA ESTIMADA=	=PROMEDIO(T)	Coeficiente de asimetría	1,87838492924543
11	TOTAL ESTIMADO=	=98*PROMEDIO(T)	Rango	4,1
12	VAR(MEDIA)=	=(1-8/98)*($D$8)/8	Mínimo	3,8
13	VAR(TOTAL)=	=98^2*$B$12	Máximo	7,9
14	ER(MEDIA)=	=100*RAIZ(B12)/B10	Suma	40,1
15	ER(TOTAL)=	=100*RAIZ(B13)/B11	Cuenta	8
16	CONFIANZA(MEDIA)=	=2*RAIZ(B12/0,05)	Nivel de confianza(95,0%)	1,07475886815483
17	CONFIANZA(TOTAL)=	=2*RAIZ(B13/0,05)		

FIGURA 2-13

	A	B	C	D
1	T		TI	
2	4,2			
3	5,1		Media	5,0125
4	7,9		Error típico	0,45451603
5	3,8		Mediana	4,85
6	5,3		Moda	5,1
7	4,6		Desviación estándar	1,28558547
8	5,1		Varianza de la muestra	1,65287057
9	4,1		Curtosis	4,24373466
10	MEDIA ESTIMADA=	5,0125	Coeficiente de asimetría	1,07030493
11	TOTAL ESTIMADO=	491,225	Rango	4,1
12	VAR(MEDIA)=	0,189720754	Mínimo	3,8
13	VAR(TOTAL)=	1822,078125	Máximo	7,9
14	ER(MEDIA)=	8,689665035	Suma	40,1
15	ER(TOTAL)=	8,689665035	Cuenta	8
16	CONFIANZA(MEDIA)=	3,896851685	Nivel de confianza(95,0%)	1,07475886
17	CONFIANZA(TOTAL)=	381,7934651		

FIGURA 2-14

el número de valores de la tabla de resultados será igual al número de muestras.

En una hoja nueva: Haga clic en esta opción para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro.

En un libro nuevo: Haga clic en esta opción para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

EJERCICIO 2-3. La Concejalía de Familia y Asuntos Sociales de un Ayuntamiento está interesada en medir la influencia de la publicidad televisiva en las familias de un municipio, y decide realizar una encuesta por muestreo para estimar el número promedio de horas por semana que se ve la televisión en los hogares del municipio. Éste comprende dos barrios, A y B, y un área rural, y se sabe que existen 155 hogares en el barrio A, 62 en el barrio B, y 93 en el área rural. La Concejalía tiene tiempo y dinero suficientes para entrevistar 40 hogares (20 del barrio A, 8 del barrio B y 12 del área rural), midiendo en cada uno el tiempo que se ve la televisión en horas por semana. Se obtienen los datos siguientes:

Pueblo A (estrato I): 35; 28; 26; 41; 43; 29; 32; 37; 36; 25; 29; 31; 39; 38; 40; 45; 28; 27; 35; 34

Pueblo B (estrato II): 27; 4; 49; 10; 15; 41; 25; 30

Área rural (estrato III): 8; 15; 21; 7; 14; 30; 20; 11; 12; 32; 34; 24

Estimar el tiempo promedio que se ve la televisión, en horas por semana, en cada uno de los estratos y en todo el municipio, fijando límites para el error de estimación a través de intervalos de confianza al 95%.

Al pulsar *Aceptar* en la Figura 2-5, se obtiene la muestra aleatoria simple de tamaño 10 con reposición de la columna C de la Figura 2-6, que ha sido extraída de la población de 22 elementos de la columna B. Si la muestra se quiere sin reposición, se utiliza este mismo procedimiento hasta obtener tantos elementos distintos como tamaño muestral se requiera.

Adicionalmente XLSTAT dispone de una utilidad para la extracción de muestras utilizando los métodos que hemos analizado en la parte teórica. Para ello seleccionaremos en Preparación de datos fi Muestreo de Datos.

Comenzamos introduciendo los datos en una hoja de cálculo de Excel. A continuación, para elegir la muestra, en el menú *Herramientas* de Excel elegimos *Análisis de datos*, seleccionamos *Muestra,* y rellenamos la pantalla de entrada como se indica en la Figura 2-8. Al pulsar *Aceptar* se obtiene la MUESTRA de tamaño 12 de la Figura 2-9. Mediante las fórmulas de la Figura 2-9 se obtienen los resultados de la Figura 2-10.

Hemos obtenido que el importe medio consumido se estima en 209,583, y que la proporción de familias cuyos dos miembros principales trabajan es del 75%. Los errores absolutos de estas estimaciones (medidos a través de sus varianzas) son 655,745 y 0,0175. En términos relativos, más fácilmente inter-

FIGURA 2-15

	Columna1		Columna2		Columna3	
3	**Media**	20,9	**Media**	20,3333333	**Media**	19
4	Error típico	1,33862208	Error típico	0,62731434	Error típico	2,70241186
5	Mediana	21,5	Mediana	15	Mediana	17,5
6	Moda	35	Moda	#N/A	Moda	#N/A
7	Desviación estándar	5,84620248	Desviación estándar	10,861943	Desviación estándar	9,36142357
8	**Varianza de la muestra**	**35,3578947**	**Varianza de la muestra**	**285**	**Varianza de la muestra**	**87,6363636**
9	**Curtosis**	**-1,0460008**	**Curtosis**	**-0,9607396**	**Curtosis**	**-1,2178641**
10	**Coeficiente de asimetría**	**0,20737524**	**Coeficiente de asimetría**	**0,6428324**	**Coeficiente de asimetría**	**0,3901415**
11	Rango	20	Rango	47	Rango	27
12	Mínimo	25	Mínimo	2	Mínimo	7
13	Máximo	45	Máximo	49	Máximo	34
14	Suma	378	Suma	183	Suma	228
15	Cuenta	20	Cuenta	9	Cuenta	12
16	**Nivel de confianza(95,0%)**	**2,78293175**	**Nivel de confianza(95,0%)**	**12,9766185**	**Nivel de confianza(95,0%)**	**5,94797159**

FIGURA 2-16

pretables, estos errores se cuantifican en el 12,2% y 17,6% respectivamente (a través de los coeficientes de variación de los estimadores).

Comenzamos introduciendo los datos como la variable *T* en una hoja de cálculo de Excel. A continuación, para calcular los estadísticos necesarios, en el menú *Herramientas* de Excel elegimos *Análisis de datos*, seleccionamos *Estadística descriptiva,* y rellenamos la pantalla de entrada como se indica en la Figura 2-11. Al pulsar *Aceptar* se obtienen los estadísticos muestrales de la Figura 2-12. Por último, se calculan los estimadores y sus errores según las fórmulas de la Figura 2-173; ésta se nos llevan a los resultados de la Figura 2-14.

	L	M	N	O	P	Q	R	S
1	Nh	nh	Wh	Wh(1-nh/Nh)	Sh2	Wh2(1-nh/Nh)Sh2/nh	median	Confianza
2	155	20	=L2/$L$5	=N2*(1-M2/L2)	=$F$6	=N2^2*O2*P2/M2	=$F$3	
3	62	8	=L3/$L$5	=N3*(1-M3/L3)	=$H=$8	=N3^2*O3*P3/M3	=$H$3	
4	93	12	=L4/$L$5	=N4*(1-M4/L4)	=$J$9	=N4^2*O4*P4/M4	=$J$3	
5	=SUMA(Nh)					=SUMA(Q2:Q4)	=PROMEDIO(mediah)	=2*RAIZ(Q5)

FIGURA 2-17

	L	M	N	O	P	Q	R	S
1	Nh	nh	Wh	Wh(1-nh/Nh)	Sh2	Wh2(Y-nh/Nh)Sh2/nh	mediah	Confianza
2	155	20	0,5	0,435483871	25,3578947	0,384844822	33,9	
3	62	8	0,2	0,174193548	285	1,241129032	20,8333333	
4	93	12	0,3	0,261290323	87,5263158	0,572903263	18	
5	310					**2,198837157**	**24,4111111**	**2,96548301**

FIGURA 2-18

Se observa que el tiempo medio por operario para terminar la tarea es de 5,0125 minutos, con un error relativo de muestreo del 9,69%. El tiempo total para terminar la tarea se estima en 491,225 minutos, con un error relativo de muestreo del 9,69%. Los errores absolutos de muestreo para estimar el tiempo medio y el total son 0,189 y 1822,078 respectivamente. El coeficiente de curtosis no está en el intervalo [-2,2], luego no podemos suponer normalidad, con lo que intervalo de confianza al 95% para la media, de anchura 1,07475886, no es válido.

Al no existir normalidad, utilizamos como intervalos de confianza:

	T	U
1	*Wh2Sh2/nh*	*confianza*
2	=N2^2*P2/M2	
3	=N3^2*P3/M3	
4	=N4^2*P4/M4	
5	**=SUMA(T2:T4)**	**=2*RAIZ(T5)**

FIGURA 2-19

	T	U
1	*Wh2Sh2/nh*	*confianza*
2	0,441973684	
3	1,425	
4	0,657272727	
5	**2,524246411**	**3,177575435**

FIGURA 2-20

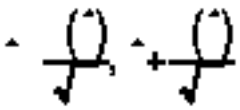
EJERCICIO 2-3. Las mil familias de una población se clasifican en tres estratos, renta baja, renta media y renta alta, para los que se conocen los datos de la tabla adjunta:

Estrato ↓	σ_i	W_i
I	4	0,6
II	12	0,3
III	80	0,1

Se pide:

Determinar el tamaño de muestra que con afijación proporcional proporciona una varianza del estimador de la media igual a 5, considerando muestreo con y sin reposición. Realizar las respectivas afijaciones proporcionales. ¿Qué resultados se obtendrían con afijación de mínima varianza? Realizar las respectivas afijaciones de mínima varianza. Comentar todos los resultados y compararlos.

Determinar también el tamaño de muestra para afijación óptima con costes C_1=1000, C_2=1200 y C_3=2000, considerando el muestreo con y sin reposición. Realizar las respectivas afijaciones óptimas. Comprobar que los resultados coinciden para costes unitarios con los de afijación de mínima varianza.



cuya anchura es 2  . Esta anchura suele considerarse como un límite para el error de estimación. Se observa que estas anchuras son mayores que con normalidad, ya que en este caso las estimaciones son menos precisas (errores mayores).

Comenzamos introduciendo los datos en tres columnas, una por cada estrato, en una hoja de cálculo de Excel. A continuación, para calcular los estadísticos necesarios en cada estrato, en el menú *Herramientas* de Excel elegimos *Análisis de datos*, seleccionamos *Estadística descriptiva,* y rellenamos la pantalla de entrada como se indica en la Figura 2-15. Al pulsar *Aceptar*, se obtienen los estadísticos muestrales

por estrato de la Figura 2-16. Se observa que el tiempo promedio que se ve la televisión en el barrio *A* es 33,9 horas por semana, en el barrio B 20,33, y en la zona rural 19. Las cuasivarianzas muestrales son 33,35, 285, y 87,63 horas por semana respectivamente en cada estrato, y al dividirlas por el tamaño muestral seleccionado en cada estrato obtenemos los errores de los estimadores en cada estrato (suponiendo muestreo con reposición): 33,35/20 = 1,667; 285/8 = 35,62 y 87,63/12 = 7,3. Como los coeficientes de asimetría y curtosis en cada estrato están en el intervalo [- 2,2], puede suponerse normalidad, con lo que los límites para el error de estimación en cada estrato (suponiendo muestreo con reposición) serán los radios de los intervalos de confianza al 95%, es decir, 2,7829, 12,97 y 5,94 respectivamente. Si el muestreo es sin reposición, las varianzas en cada estrato hay que multiplicarlas por $(1 - n_h/N_h)$ $h = 1, 2, 3$.

Para hallar la estimación del tiempo promedio que se ve la televisión en todo el municipio en horas por semana, y su error para muestreo sin reposición, se tendrán en cuenta las siguientes expresiones:

$$\hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi} = \sum_{h=1}^{L} W_h \bar{x}_h \qquad \hat{V}(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - f_h\right) \frac{\hat{S}_h^2}{n_h}$$

La Figura 2-17 presenta las fórmulas para el cálculo del estimador de la media estratificada para todo el municipio, su error de muestreo, y el radio del intervalo de confianza al 95%. La Figura 2-18 presenta los resultados.

La estimación del tiempo promedio que se ve la televisión en todo el municipio en horas por semana en muestreo con reposición es la misma que sin reposición, y su error de muestreo se calcula mediante la siguiente expresión:

$$\hat{V}(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\hat{S}_h^2}{n_h}$$

La Figura 2-19 presenta las fórmulas, y la Figura 2-20 presenta los resultados.

Como es habitual en los problemas de muestreo estratificado, comenzamos recopilando los datos necesarios para el problema.

$W_1=0,6=N_1/N \quad N_1=600 \qquad \sigma_1^2=16=(N_1-1)S_1^2/N_1 \quad S_1^2=6,02 \quad S_1=4,003$

$W_2=0,3=N_2/N \quad N_2=300 \qquad \sigma_2^2=144=(N_2-1)S_2^2/N_2 \quad S_2^2=144,5 \quad S_2=12,02$

$W_3=0,1=N_3/N \quad N_3=100 \qquad \sigma_3^2=6400=(N_3-1)S_3^2/N_3 \quad S_3^2=6464,6 \quad S_3=80,4$

Tenemos entonces:

Estrato ↓	S_h	S_h^2	σ_h	σ_h^2	N_h	W_h
I	4,003	6,02	4	16	600	0,6
II	12,02	144,5	12	144	300	0,3
III	80,4	6464,6	80	6400	100	0,1

Afijación proporcional sin reposición

$$e^2 = V\left(\hat{\bar{X}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\sum_{h=1}^{L} W_h S_h^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} \cong 122$$

Una vez hallado el tamaño de muestra, realizamos la afijación como sigue:

$$n_h = kN_h \text{ con } k = \frac{n}{N} = \frac{122}{1000} = 0{,}122 \Rightarrow \begin{cases} n_1 = kN_1 = 0{,}122 \cdot 600 \cong 73 \\ n_2 = kN_2 = 0{,}122 \cdot 300 \cong 37 \\ n_3 = kN_3 = 0{,}122 \cdot 100 \cong 12 \end{cases}$$

Afijación proporcional con reposición

$$e^2 = V\left(\hat{\bar{X}}\right) = \frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2 \Rightarrow n = \frac{\sum_{h=1}^{L} W_h \sigma_h^2}{e^2} \cong 139$$

Se observa que el tamaño muestral necesario para cometer el mismo error que sin reposición es ahora superior. Ello es debido a que el muestreo con reposición es menos preciso que el muestreo sin reposición. Una vez hallado el tamaño de muestra realizamos la afijación proporcional como sigue:

$$n_h = kN_h \text{ con } k = \frac{n}{N} = \frac{139}{1000} = 0{,}139 \Rightarrow \begin{cases} n_1 = kN_1 = 0{,}139 \cdot 600 \cong 83 \\ n_2 = kN_2 = 0{,}139 \cdot 300 \cong 42 \\ n_3 = kN_3 = 0{,}139 \cdot 100 \cong 14 \end{cases}$$

Afijación de mínima varianza sin reposición

$$e^2 = V\left(\hat{\bar{X}}\right) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} = 35$$

Una vez hallado el tamaño de muestra, realizamos la afijación de mínima varianza como sigue:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \Rightarrow \begin{cases} n_1 = 35 \cdot \dfrac{N_1 S_1}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 6 \\ n_2 = 35 \cdot \dfrac{N_2 S_2}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 9 \\ n_3 = 35 \cdot \dfrac{N_3 S_3}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 20 \end{cases}$$

Afijación de mínima varianza con reposición

$$e^2 = V\left(\hat{\bar{X}}\right) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{e^2} \cong 40$$

Se observa que el tamaño muestral necesario para cometer el mismo error que sin reposición es ahora superior. Una vez hallado el tamaño de muestra realizamos la afijación de mínima varianza como sigue:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} \Rightarrow \begin{cases} n_1 = 35 \cdot \dfrac{N_1 \sigma_1}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 7 \\ n_2 = 35 \cdot \dfrac{N_2 \sigma_2}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 10 \\ n_3 = 35 \cdot \dfrac{N_3 \sigma_3}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 23 \end{cases}$$

Afijación óptima sin reposición

EJERCICIO 2-4. Para tomar la decisión de mantener una determinada prestación social, se pretende tomar una muestra aleatoria simple entre las 1.250 potenciales familias beneficiarias de un municipio y enviarles un cuestionario a través del cual manifiesten si son favorables a la política existente o preferirían destinar el dinero a otro fin.

¿Cuál deberá ser el número apropiado de familias a muestrear de entre las 1.250 para obtener una estimación sobre la proporción de familias favorables al mantenimiento de la política con un error de muestreo inferior al 12% y una confianza del 90%?

Si de la encuesta realizada el año anterior se sabe que la proporción de familias favorables a la política estará entre el 75% y el 85%, ¿cuál debería ser en este caso el número apropiado de familias encuestadas del apartado anterior?

Si finalmente se decidió enviar cuestionarios a 100 familias, de los cuales tan sólo 35 no se manifestaron favorables a la prestación social existente, estimar la proporción del número apropiado de familias encuestadas de entre los 1250 para obtener una estimación.

$$V(\bar{x}_{st}) = e^2 = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)$$

$$-\frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} \cong 35$$

Una vez hallado el tamaño de muestra, realizamos la afijación óptima como sigue:

$$n_h = n \cdot \frac{N_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{C_h}} \Rightarrow \begin{cases} n_1 = 35 \cdot \dfrac{N_1 S_1 / \sqrt{C_1}}{N_1 S_1 / \sqrt{C_1} + N_2 S_2 / \sqrt{C_2} + N_3 S_3 / \sqrt{C_3}} \cong 7 \\ n_2 = 35 \cdot \dfrac{N_2 S_2}{N_1 S_1 / \sqrt{C_1} + N_2 S_2 / \sqrt{C_2} + N_3 S_3 / \sqrt{C_3}} \cong 10 \\ n_3 = 35 \cdot \dfrac{N_3 S_3}{N_1 S_1 / \sqrt{C_1} + N_2 S_2 / \sqrt{C_2} + N_3 S_3 / \sqrt{C_3}} \cong 18 \end{cases}$$

Afijación óptima con reposición

$$V(\bar{x}_{st}) = e^2 = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h \sigma_h \sqrt{c_h}\right) \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h \sigma_h \sqrt{c_h}\right)}{e^2} = 40$$

Se observa que el tamaño muestral necesario para cometer el mismo error que sin reposición es ahora superior. Una vez hallado el tamaño de muestra realizamos la afijación óptima como sigue:

$$n_h = n \cdot \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h \sigma_h / \sqrt{c_h}} \Rightarrow \begin{cases} n_1 = 40 \cdot \dfrac{N_1 \sigma_1 / \sqrt{c_1}}{N_1 \sigma_1 / \sqrt{c_1} + N_2 \sigma_2 / \sqrt{c_2} + N_3 \sigma_3 / \sqrt{c_3}} \simeq 8 \\ n_2 = 40 \cdot \dfrac{N_2 \sigma_2}{N_1 \sigma_1 / \sqrt{c_1} + N_2 \sigma_2 / \sqrt{c_2} + N_3 \sigma_3 / \sqrt{c_3}} \simeq 12 \\ n_3 = 40 \cdot \dfrac{N_3 \sigma_3}{N_1 \sigma_1 / \sqrt{c_1} + N_2 \sigma_2 / \sqrt{c_2} + N_3 \sigma_3 / \sqrt{c_3}} \simeq 20 \end{cases}$$

Si utilizamos costes unitarios los cálculos son exactamente los mismos que para la afijación de mínima varianza, luego los resultados también lo son. Se observa que tanto en muestreo con reposición como sin reposición la afijación que menos tamaño muestral necesita para cometer un determinado error de muestreo es la afijación de mínima varianza, y en este caso también la óptima.

CAPÍTULO III
EXPLORACIÓN DE LOS DATOS

AURELIA VALIÑO CASTRO

3.1. EXPLORACIÓN DE DATOS: UN PRIMER PASO IMPRESCINDIBLE

Una vez que tenemos los datos de una muestra estadística nos urge empezar a estudiar lo que nos motivó a su recolección: ¿son las familias numerosas del país A más pobres que las del país B? o ¿cuán eficiente es la educación pública para las familias inmigrantes? o ¿en qué medida las políticas a favor de la familia permiten elevar la tasa de natalidad de un país?... o cualquier otra igual o más interesante. Pero para poder desarrollar las respuestas a las que ayudarán los capítulos siguientes de este libro, se hace necesario un análisis previo de los datos. La mayoría de las técnicas que se emplean para dar respuesta a esas preguntas exigen unas características y unas condiciones en los datos que es necesario verificar antes de aplicarlas, como por ejemplo: condiciones de normalidad. En otras ocasiones conocer los datos nos sirve para entender mejor los resultados que podamos obtener con técnicas relativamente complejas. Incluso son útiles por sí solos para conocer mejor la muestra con la que trabajamos y ayudarnos a plantearnos posibles preguntas que en otro caso no se nos hubieran ocurrido, dirigiendo así la investigación de una forma más eficiente.

Así pues, es imprescindible conocer individualmente las variables con las que vamos a operar, si existe alguna relación entre ellas, las medidas estadísticas que las describen, comprobando si hay alguna agrupación de los datos, si están concentrados o dispersos, las diferencias entre los grupos, identificar tendencias, ver si hay datos anómalos o ausencia de algún dato, o simplemente conocer datos tan básicos, pero imprescindibles, como el valor mínimo, máximo o medio de los datos de una variable.

En esta primera aproximación a los datos, el análisis gráfico es un medio visual muy útil que ofrece información «a primera vista», aunque la información que proporcionen los gráficos no puede sustituir a las medidas de diagnóstico estadístico.

Hoy en día todos los paquetes estadísticos informáticos traen incluidos programas que permiten realizar de forma automática una exploración previa de los datos. El análisis que vamos a realizar en este capítulo se puede seguir a través de dos conjuntos de herramientas que se derivan del pro-

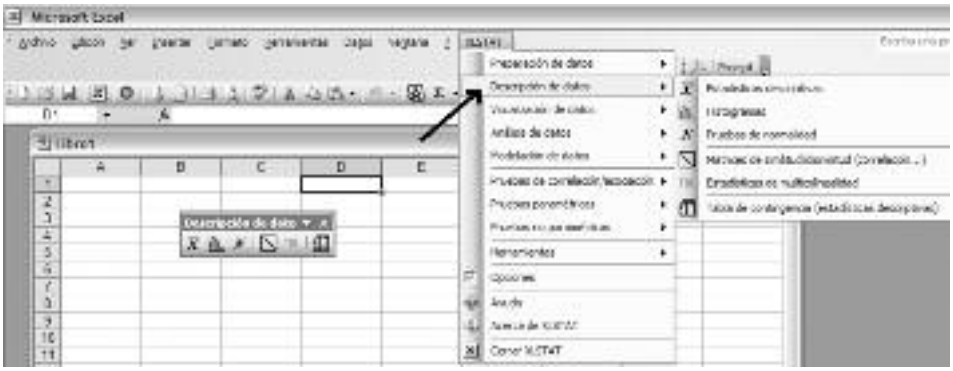


FIGURA 3-1

grama una vez que se ha incorporado a Excel: descripción de datos y visualización de datos. Haciendo "clic" sobre XLSTAT se despliegan los contenidos de éste y situando el cursor sobre las pestañas de «descripción de datos» y «visualización de datos» se despliegan los contenidos de las mismas, tal y como se muestra en las figuras 3.1 y 3.2 siguientes.

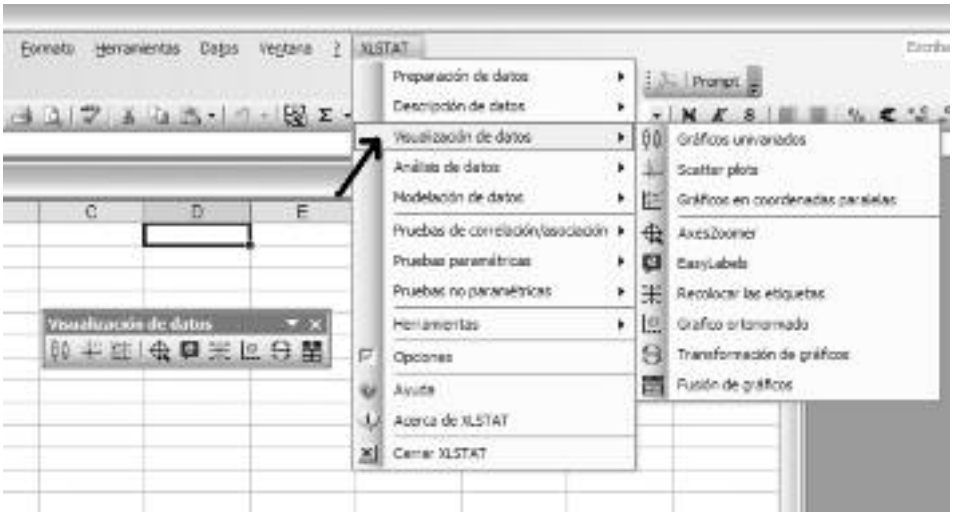


FIGURA 3-2

Mientras realizamos los análisis de exploración de datos puede interesarnos tener visible el contenido de las pestañas de descripción o visualización. Para ello en el menú de XLSTAT nos situamos sobre «Herramien-

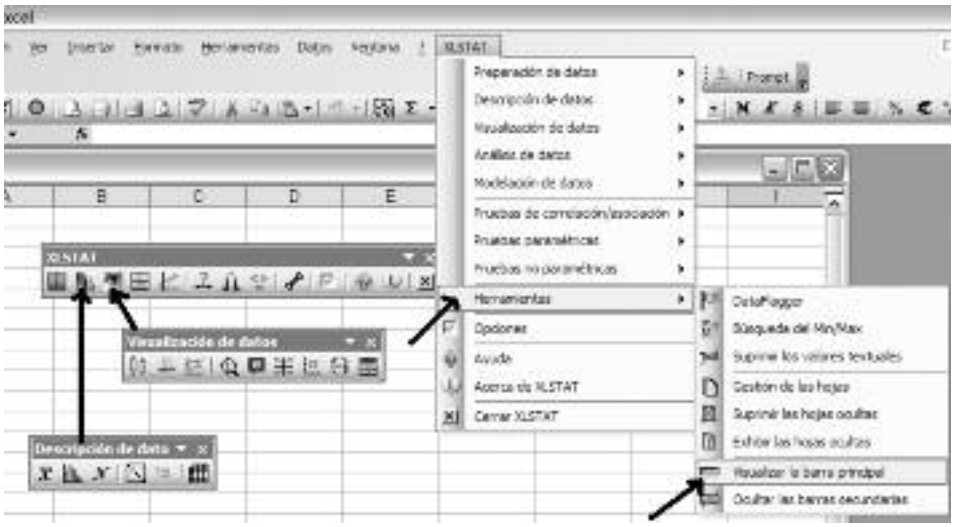


FIGURA 3-3

tas» y en su despliegue sobre «visualizar la barra principal». Una vez que se ha abierto esta última, pinchando en el segundo gráfico de la izquierda obtenemos la barra secundaria de descripción de datos; y en la representación de un ojo, la barra secundaria de visualización de datos. Si nos estorba la visión de la barra principal, se puede cerrar dejando abiertas una o varias secundarias, que se pueden cerrar a su vez una a una pinchando en la "x" de cierre o todas a la vez en la opción de «ocultar barras secundarias». Todo este proceso se muestra en la figura 3-3.

3.2. ESTADÍSTICAS DESCRIPTIVAS

Bajo la expresión de estadísticas descriptivas se encuentran todo un conjunto de análisis que se realizan sobre los datos disponibles. El tipo de análisis depende a su vez del tipo de datos. En nuestro análisis de un grupo social: la familia, los datos se refieren a características de las mismas, ya sean económicas (renta, salarios, subsidios o transferencias públicas recibidas…), sociales (personas que conviven en un hogar, nivel de educación del sustentador principal, malos tratos…), demográficas (número de hijos, fallecimientos, nacimientos, edad de los componentes…) religiosas (filiación religiosa de sus miembros, grado de actividad religiosa…), legales (estado civil, número de adopciones, abusos declarados…), etc. Cada una de estas características constituye una *variable* compuesta por una determinada cantidad de datos (denominados observaciones o casos), cuya amplitud depende de la muestra tomada, según las especificaciones del capítulo anterior dedicado al muestreo y sus técnicas.

Los datos pueden ser de carácter cuantitativo (p.e. 3.000 euros) y las variables se denominan *cuantitativas* (p.e renta mensual de la familia), o de carácter cualitativo (p.e. soltero, casado, viudo, divorciado) y las variables se denominan *cualitativas* o *categóricas* (p.e. estado civil del sustentador principal).

Así pues, las *variables cuantitativas* son las que expresan una cantidad y pueden medirse o expresarse numéricamente. Pueden ser *continuas*, si pueden tomar todos los valores dentro de un rango numérico determinado, como la edad (p.e 5 años, tres meses y dos días), la renta (p.e 3.525,33) o el peso (p.e. 65,3 kg); o *discretas*, si toman un valor determinado y concreto (se expresan con números enteros), como el número de hijos, número de casados, número de mayores de 65 años, número de nacidos, etc..

Las *variables cualitativas*, también llamadas *categóricas*, miden una cualidad o atributo. Las observaciones no pueden ser cuantificadas, pero pueden asociarse a un número o letra para ser clasificadas en categorías. Por ejemplo, la variable sexo tiene dos categorías: hombre o mujer, que se pueden asociar con V y H o, lo que es más corriente: 0 y 1. Este tipo de variables con dos categorías se llaman *dicotonómicas o binarias* y son muy frecuentes; otros ejemplos: fumador o no fumador, nacional o extranjero, jubilado o activo; etc.. Pero en ocasiones son necesarias más categorías, como por ejemplo: estado civil (soltero, casado, viudo, divorciado), composición familiar (familias sin hijos, con 1 hijo, con 2, con 3 o más), situación laboral del padre (ocupado, parado, con incapacidad total, jubilado), etc. Todos los ejemplos anteriores coinciden con lo que se denomina escalas nominales; serían *variables categóricas nominales*. Son variables cuyas observaciones se ajustan por categorías que no imponen ningún orden «natural» o relación entre sí. Cuando, por el contrario, existe cierto orden «natural» o jerarquía en las escalas de las categorías, nos encontramos con *variables categóricas ordinales*. Un ejemplo de estas últimas serían los resultados de un examen (suspenso o insuficiente, aprobado o suficiente, notable, sobresaliente) o nivel de estudios del padre (analfabeto, estudios primarios, estudios básicos, estudios medios, estudios superiores).

Cada una de las variables anteriores requiere un análisis estadístico específico. Así, sobre las variables cuantitativas se realizan estudios de concentración o dispersión. *Medidas de concentración* o de tendencia central son la media, la mediana o la moda; las de *dispersión*: la varianza, la desviación típica, la desviación estándar, el coeficiente de variación y de *posición*: cuartiles y percentiles. Y sobre las variables cualitativas o categóricas se realizan estudios de frecuencias (aunque también se realizan para las variables cuantitativas), moda, peso de la modalidad.

A continuación desarrollamos estos análisis.

Medidas de concentración, centralización o de tendencia central

La mejor forma de operar con un conjunto de datos sin perder información es reducirlos a un único valor. Se intenta buscar el valor más representativo del conjunto de datos que se analizan. Esto es lo que pretenden las medidas de concentración a través de medir la localización central de los datos en una muestra. A esta operación, sólo aparentemente compleja, ayudan la media, la mediana y la moda.

- *Media:* Es el valor aritmético medio de un conjunto de datos. Es la suma de los valores de los datos, eventualmente ponderada, dividida por el número de valores utilizados, o por la suma de los pesos si los datos son ponderados. Si (X1, X2,...,Xn) son los n datos que tenemos recogidos de la variable en cuestión, el valor medio vendrá dado por $\frac{1}{n}\sum_{i=1}^{n} x_i$. En el caso de que los datos sean ponderados y la suma de los valores ponderados es $S=\sum_{i=1}^{n} w_i x_i$, y la suma de ponderaciones es *Sw,* entonces la media es *m =S/Sw.*

- *Mediana:* es la valoración equidistante de los extremos. Es el valor que deja a la mitad de los datos por encima del mismo y a la otra mitad por debajo, o el valor del medio cuando los valores se ordenan de menor a mayor. Si la media y la mediana son iguales, la distribución de la variable es simétrica. La media es muy sensible a los valores extremos. Sin embargo, la mediana es menos sensible.

- *Moda:* es la observación que presenta mayor frecuencia o que se presenta más frecuentemente.

- En una *distribución normal* la media, mediana y moda tienen igual valor.

Medidas de dispersión

El análisis de los datos cuantitativos precisa de más medidas que las de concentración para lograr un resumen de la información que presenta. Es necesario también observar si los datos están más o menos dispersos.

- *Varianza*. Es la media de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de la distribución. Tiene utilidad para calcular otras medidas de dispersión como la desvia-

ción estándar o típica (en el denominador se pone n-1 cuando se quieren aproximar los valores poblacionales).

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Cuando tenemos datos ponderados: $s(n)^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{S_W}$

O $s(n-1)^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{S_W - S_W/n}$

- *Desviación típica o estándar*, da una medida de la concentración o dispersión de los datos respecto a la media, si están concentrados la desviación es menor y si están dispersos la desviación es mayor. Se ve muy afectada por los valores extremos o outliers. Por ello se puede considerar también un indicador de la presencia de estos valores. Es muy útil como medida de dispersión para dos muestras que tienen medias muy similares; en este caso la muestra con menor desviación típica tiene los datos mas concentrados en torno a la media y cualquier dato de la muestra que se tome al azar tendrá más probabilidad de estar cerca de la media. La desviación típica nunca da un valor negativo. Si todos los valores de la muestra son iguales, coinciden con la media y la desviación típica es cero. Se define como la raíz cuadrada de la varianza s(n)= $\sqrt{s^2}$

- *Coeficiente de variación*, mide la dispersión relativa obtenida dividiendo la desviación típica por la media. Este coeficiente permite comparar la dispersión de variables cuyas unidades son diferentes (diferentes escalas), o que tienen medias muy diferentes, s(n)/m

Medidas de localización o posición

Cuando tenemos distribuciones simétricas y no hay valores extremos, los valores de la media y la mediana están próximos y la media y la desviación típica son las medidas, más representativas y adecuadas de los datos.

Pero cuando las distribuciones son asimétricas es más adecuado acudir a otras medidas. En este caso se utilizan la mediana, el rango o amplitud y los cuartiles y percentiles.

- *Rango o amplitud*: es la diferencia entre la mayor y la menor de las observaciones de una muestra. Su principal problema es que no mide como se distribuyen los datos entre estos valores y que es totalmente inapropiado cuando hay valores extremos. Por estos problemas es una medida de apoyo, pero nunca una medida útil por sí sola.

- *Percentiles*: es el valor de la observación que indica el porcentaje de la muestra que es igual o menor a ese valor; por ejemplo, el percentil 60 es el valor de la observación que deja por debajo al 60 por cien de la muestra.

- *Cuartil*: se divide la muestra en cuatro partes y cada una es un cuartil. El primer cuartil (Q_1) es el valor de la observación que deja por debajo el 25 por cien de la muestra; el segundo cuartil (Q_2), el 50 por cien de la muestra (coincide con la mediana) y el tercer cuartil (Q_3), el 75% de la muestra.

- *Rango intercuartil*: la diferencia entre el menor y el mayor cuartil (Q_3–Q_1), este valor también informa de la dispersión de una muestra. El rango intercuartil abarca el 50 por cien de la muestra, dejando el 25 por cien por arriba y el 25 por cien por abajo, por lo tanto elimina la influencia de los valores extremos.

Distribución de frecuencias

En el análisis exploratorio de los datos reviste una gran importancia la distribución de frecuencias, también llamada tabla de frecuencias. En esta tabla se asocia a cada dato o subgrupo de datos (intervalo de clase o clase) el número de observaciones que coinciden con su valor o grupo de valores (frecuencia). Es decir: frecuencia es el número de veces que se repite una observación en la muestra y el conjunto de valores que ha tomado una variable junto con sus frecuencias es lo que se denomina *distribución de frecuencias*.

En un conjunto de datos u observaciones de una muestra lo más frecuente es que no exista un único valor constante para cada observación y que los datos estén dispersos. Por ello es necesario ver si hay algún patrón o pauta de comportamiento en las observaciones. La distribución de frecuencias ayuda a ver como se distribuye la variable y su forma: si es simétrica o asimétrica, si tiene picos, cómo son y cuántos, y también a ver cual

es la dispersión respecto del promedio, o relaciones entre las medidas de dispersión, concentración y posición.

Los distintos tipos de frecuencias que nos ayudan en este análisis son:

- *Frecuencia absoluta*: es el número de veces que se repite un valor (x_i) de una variable (X) en la muestra. Se denota por n_i. Y cumple $\sum_{1}^{k} n_i = n_1 + \ldots + n_k = N$. Está influida por el tamaño de la muestra, al aumentar el tamaño de la muestra aumentará también el tamaño de la frecuencia absoluta. Esto hace que no sea una medida útil para poder comparar. Para ello se prefiere la siguiente medida.

- *Frecuencia relativa*: es el cociente entre la frecuencia absoluta y el tamaño de la muestra (N). La denotaremos por $f_i = n_i / N$. Y cumple

$$\sum_{1}^{k} f_i = 1$$

- *Frecuencia absoluta acumulada*: La frecuencia absoluta acumulada de un valor (x_i) de la variable, es el número de veces que ha aparecido en la muestra un valor menor o igual que el de la variable; es decir, es la suma de las frecuencias absolutas de los valores de la variable anteriores o iguales a su valor (x_i) y lo representaremos por $N_i = \sum_{1}^{i} n_i$.

- *Frecuencia relativa acumulada*: es la frecuencia absoluta acumulada dividida por el número total de valores de la variable. Su valor es $F_i = N_i / N$.

- *Porcentaje*: La frecuencia relativa es un tanto por uno, sin embargo, hoy día es bastante frecuente hablar siempre en términos de tantos por ciento o porcentajes, por lo que esta medida resulta de multiplicar la frecuencia relativa por 100. La denotaremos por $p_i = f_i 100$.

- *Porcentaje acumulado*. se define como la frecuencia relativa acumulada por 100 $P_i = F_i 100$.

Medidas de forma o simetría

Estas medidas tienen en cuenta la forma de la distribución. Se trata de ver qué información aporta teniendo en cuenta la forma que adopta la distribución de los datos: si la distribución es simétrica o no, si es achatada o presenta un apuntamiento. Estas medidas, precisamente por ser medidas de la forma que presenta la distribución de los datos, tienen mucho que ver con el análisis gráfico.

- *Asimetría*: La simetría es importante para saber si los valores de la variable se concentran en una determinada zona del recorrido de la misma. Se dice que una distribución de frecuencias es *simétrica* cuando los valores de la variable son equidistantes de un valor central dos a dos y cada par de valores equidistantes tienen la misma frecuencia. Así pues, la simetría se produce cuando hay igualdad, pero hay que especificar respecto a qué. Una posibilidad es respecto a la mediana. Si la variable es discreta, lo será respecto a la media. En variables continuas coinciden la media y la mediana. Si la variable es continua y unimodal, es simétrica si coinciden la media, la moda y la mediana. En el caso de que las frecuencias más altas se encuentren en el lado izquierdo de la media, y en el derecho haya frecuencias más pequeñas o cola, diremos que hay *asimetría negativa*. Si por el contrario las frecuencias más altas se dan en el lado derecho de la media y las más bajas en el izquierdo, diremos que hay *asimetría positiva*.

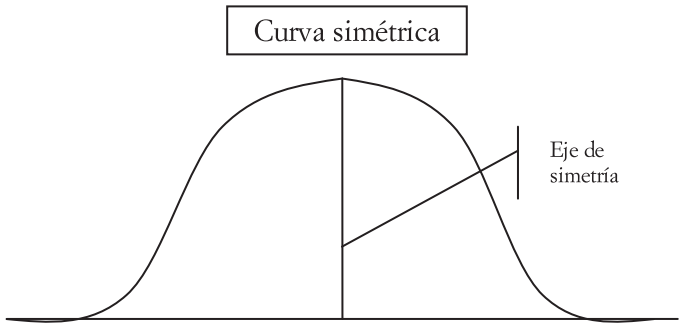


FIGURA 3-2

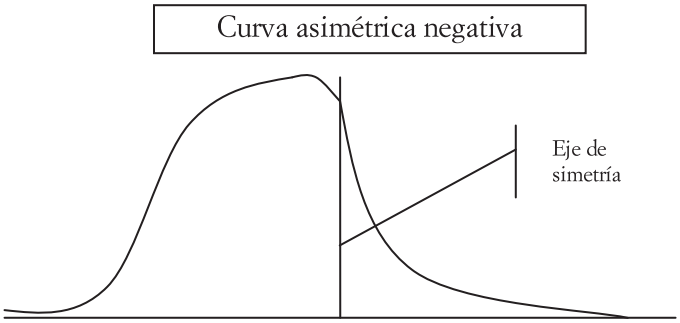


FIGURA 3-3

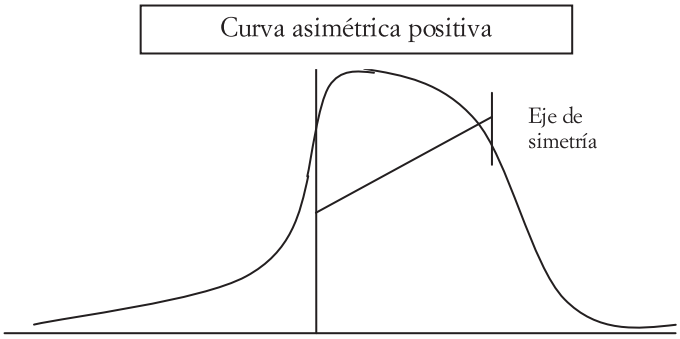


FIGURA 3-4

No siempre es fácil distinguir la simetría de una distribución a través del análisis gráfico, por ello se utilizan índices que se apoyan en otras medidas, como por ejemplo:

– Indicador de Yule-Bowley: utiliza los cuartiles para comprobar la simetría. Así, una distribución es simétrica si, tomando como referencia la mediana(Q_2) el cuartil de la izquierda (Q_1) es igual al cuartil de la derecha (Q_3); es decir, la observación que deja por debajo un cuarto de la muestra es igual a la observación que deja por debajo las tres cuartas partes de la distribución $Q_3 - Q_2 = Q_2 - Q_1$. En el caso de asimetría positiva $Q_3 - Q_2 > Q_2 - Q_1$, y en el caso de asimetría negativa $Q_3 - Q_2 < Q_2 - Q_1$. Este índice puede verse afectado por la escala de la distribución, para evitarlo se utilizaría: $A_y = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$ Se puede utilizar también como medida que no se vea afectada por cambios de referencia y escala: $-1 \leq A_y = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \leq 1$

– Coeficiente de asimetría de Pearson: la media menos la moda divididas entre la varianza. Si el valor es cero, la distribución es simétrica; si la diferencia es positiva, hay asimetría a la derecha; si la diferencia es negativa hay asimetría a la izquierda. Esta medida presenta como ventaja la facilidad de cálculo, pero el inconveniente de no ser muy precisa. De hecho sólo es válido para distribuciones con forma de campana y unimodales. El segundo coeficiente de Pearson divide el triple de la diferencia de la media y la mediana entre la desviación típica.

– Coeficiente de Asimetría de Fisher:

$$g_1 = \frac{(1/n)\sum (x_i - \bar{x})^3 n_i}{((1/n)\sum (x_i - \bar{x})^2 n_i)^{3/2}}$$

$$g_1 = \frac{m_3}{S_x^3} = \frac{\sum (x_i - \bar{x})^3 n_i / N}{S_x^3}$$

- $g_1 = 0$ (distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media)
- $g_1 > 0$ (distribución asimétrica positiva; existe mayor concentración de valores a la derecha de la media que a su izquierda)
- $g_1 < 0$ (distribución asimétrica negativa; existe mayor concentración de valores a la izquierda de la media que a su derecha)

- *Curtosis o apuntamiento:* este coeficiente mide la mayor o menor agrupación de datos en torno a la moda de una distribución. Es decir: mide la forma picuda o achatada de la misma. Se definen tres tipos de distribuciones según su grado de curtosis:

  – *Distribución mesocúrtica*: presenta un grado de concentración medio alrededor de los valores centrales de la variable. Esta forma es la que se toma como referencia, coincide con la forma acampanada de la distribución normal (campana de Gauss).

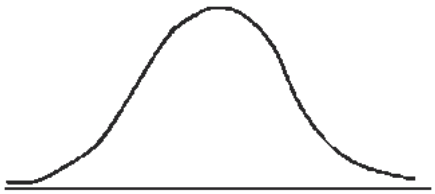


FIGURA 3-5

– *Distribución leptocúrtica*: presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

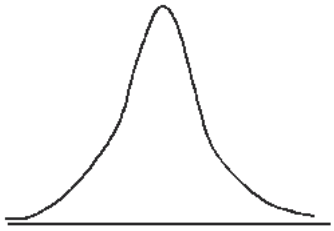


FIGURA 3-6

– *Distribución platicúrtica*: presenta un reducido grado de concentración alrededor de los valores centrales de la variable.

FIGURA 3-7

- El coeficiente de curtosis de Pearson está definido por:

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i}{n m^4} - 3$$

Si $g_2 > 0$ la distribución será leptocúrtica o apuntada

Si $g_2 = 0$ la distribución será mesocúrtica o normal

Si $g_2 < 0$ la distribución será platicúrtica o menos apuntada que lo normal.

- El coeficiente de curtosis de Fisher se define por:

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = 3$$

mesocúrtica (tan aplanada como una normal, $\gamma_2 = 0$), leptocúrtica (más apuntada que una normal, $\gamma_2 > 0$) o platicúrtica (más aplanada que una normal, $\gamma_2 < 0$).

3.3. ANÁLISIS GRÁFICO

El análisis gráfico presenta la ventaja de «visualizar» muchas de las medidas que antes hemos analizado de forma conjunta, por lo que de «un vistazo» podemos ver la síntesis de los datos que es necesaria para poder abordar cualquier análisis un poco más complejo. Por ello en muchos casos es más útil presentar la información en un gráfico que en una tabla.

Histogramas

El histograma es una representación gráfica de una variable cuantitativa en forma de barras, en la que la superficie de cada barra es proporcional a la frecuencia de los valores presentados. En el eje vertical se representan las frecuencias y en el horizontal los valores, recogiendo la «marca de clase» o la mitad del intervalo en el que están agrupados. Un histograma tiene la apariencia de un gráfico de barras vertical, pero no es lo mismo. En un histograma la frecuencia se mide por el área en la columna, mientras que en un gráfico de barras vertical la frecuencia se mide por la altura de la barra. Cuando las variables son continuas, las barras están pegadas unas a otras; cuando las variables son discretas, las barras están separadas.

La gran ventaja del histograma es que permite visualizar la pauta de variación de los datos, viendo si la distribución de los datos tiene forma de campana, si tiene o no forma de pico, si es sesgada y hacia donde, si es truncada, si tiene mas de un pico, si tiene valores extremos o outliers, etc. También nos permite comparar la forma de la distribución obtenida con otros datos preestablecidos, ya sea por el analista o con distribuciones teóricas, como la normal, la beta, la binomial, la exponencial, la lognormal, la de Pareto, etc. A efectos de la comparación con una distribución teórica se recomienda utilizar la distribución empírica acumulada.

Para la realización de un histograma se siguen los siguientes pasos:

- En primer lugar hay que determinar el rango de los datos (la diferencia entre el dato mayor y el menor).

- En segundo lugar hay que determinar el número de clases o barras. Existen distintas formas de calcular el número de clases, el método

mas tradicional es usar clases definidas por intervalos de igual anchura o amplitud, siendo el límite del menor intervalo determinado por el valor mínimo o un valor ligeramente menor del mínimo. Para evitar que algún dato coincida con el extremo de un intervalo se suele añadir un decimal a los que tengan los datos. Un criterio utilizado para establecer el número de clases es aproximadamente la raíz cuadrada del número de datos. Otro criterio es la regla de Sturgess, según la cual el número de clases es igual a $1+3{,}3 \log_{10}$ (tamaño del colectivo).

- En tercer lugar, determinar la longitud de clase, que es igual al rango entre el número de clases.

- En cuarto lugar, construir los intervalos de clases que se obtienen dividiendo el rango de los datos por el número de clases.

- Por último se realiza el histograma, teniendo en cuenta que las bases de las barras son los intervalos de clases y la altura de las barras es la frecuencia de las clases, o el número de datos incluidos en cada clase. Uniendo los puntos medios de la parte superior de las barras se obtiene el *polígono de frecuencias*. Los polígonos de frecuencias suavizan los cambios o saltos de una barra a otra en los histogramas, y son útiles para demostrar la continuidad en la variable.

En las líneas que siguen se resumen las pautas que pueden presentarse en los histogramas:

- Distribución en forma de campana. Esta distribución simétrica con un pico es la que se considera natural y por ello se la denomina *normal*. Cuando no se produce es probable que indique la existencia de algún problema que habrá que comprobar analizando con más detalle los datos.

- Distribución con doble pico: cuando aparece un valle en el centro de la distribución y picos a los lados. Puede indicar la existencia de dos distribuciones y la necesidad de separarlas. Para hacerlo será necesario estratificar los datos, o dividirlos en grupos o categorías, de forma de que los datos pertenecientes a cada grupo tengan características comunes a la categoría.

- Distribución plana: cuando no aparece ningún pico y hay dos ligeras colas a los lados. Estas distribuciones suelen ocultar varias distribuciones de campana con sus centros distribuidos a los largo de todo el recorrido de los datos. Habrá que identificar también estos procesos estratificando los datos.

- Distribución en peine: en este caso aparecen, alternándose, valores altos y bajos. Esta pauta suele reflejar errores de medición, errores en la forma de agrupar los datos, o sesgos de redondeo.

- Distribución con un pico aislado: se trata de una distribución con una forma de campana, pero que hacia una de las colas presenta un pico aislado. El suceso con el pico pequeño puede indicar una anormalidad. Estos picos unidos a distribuciones sesgadas o truncadas indican falta de eficacia en la eliminación de sucesos defectuosos. Un caso particular es cuando tenemos una distribución normal con un pico en un extremo, se presenta cuando la cola se ha cortado y acumulado en una sola categoría en el extremo del recorrido de los datos. Esto pude ocurrir cuando se han registrado mal los datos o hay un sesgo en los mismos.

- Distribución sesgada o truncada: tiene una forma asimétrica, con un pico no situado en el centro

Diagrama de Sectores

Las distribuciones de frecuencias de variables cualitativas pueden ser representadas también por gráficos de barras, representando en el eje de ordenadas las modalidades y en abscisas las frecuencias absolutas, o bien, las frecuencias relativas. Si, mediante el gráfico, se intenta comparar varias poblaciones entre sí, y los tamaños de las poblaciones son diferentes, es conveniente utilizar las frecuencias relativas, ya que en otro caso sería difícil lograr una interpretación adecuada.

Pero los diagramas típicos para las variables cualitativas son los diagramas de sectores, también conocidos como diagramas de «tartas», se divide un círculo en tantas porciones como clases tenga la variable, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa. La información que debe mostrar cada sector se refiere al número de casos dentro de cada categoría y al porcentaje que representan respecto al total. También se pueden comparar dos poblaciones, como en el caso anterior, usando para cada una de ellas un diagrama semicircular. Si los tamaños de las poblaciones son $n_1 \leq n_2$, la población más pequeña se representa con un semicírculo de radio r_1 y la mayor con radio r_2. La relación entre los radios resulta de suponer que la relación entre las áreas de las circunferencias es igual a la de los tamaños de las poblaciones respectivas; es decir: $\frac{\pi r_2^2}{\pi r_1^2} = \frac{n_2}{n_1} \Leftrightarrow r_2 = r_1\sqrt{\frac{n_2}{n_1}}$

Gráfico de caja y bigotes

O simplemente gráfico de caja. Fue inventado por John Tukey. Se trata de una representación de una muestra de datos cuantitativos, en la que se recogen los cuartiles (caja) y el recorrido de la misma (bigotes). La caja se delimita por los cuartiles mayor y menor, dividiendo el cuartil de en medio (mediana) la caja en dos; así pues, la caja representa el rango intercuartil y recoge el cincuenta por ciento de los datos. El valor de la media también se suele representar por un símbolo dentro de la caja. Y los bigotes recogen cada uno el veinticinco por ciento de los datos y se delimitan por los mayores y menores valores de la variable. Obviamente, en el caso de que tengamos valores extremos (excesivamente altos o bajos respecto al resto —outliers—) deberían excluirse para formar los bigotes, pero representarlos individualmente en el gráfico. En principio se considera que cualquier valor por encima o por debajo de 1,5 veces el rango intercuartil es un valor extremo o «outlier». La representación puede ser horizontal, como en el ejemplo siguiente, o vertical.

Además de lo expuesto, el gráfico de caja y bigotes permite hacer comparaciones para distintos conjuntos de datos representando en el mismo gráfico su diagrama de caja y bigotes. Precisamente una de las utilidades de este gráfico es que permite hacer comparaciones, pero además permite resumir la muestra cuando sus datos son muy amplios y comprobar si la distribución es sesgada y hay valores extremos. Nos da una medida de la dispersión de los datos mucho mas clara incluso que el valor de la desviación típica.

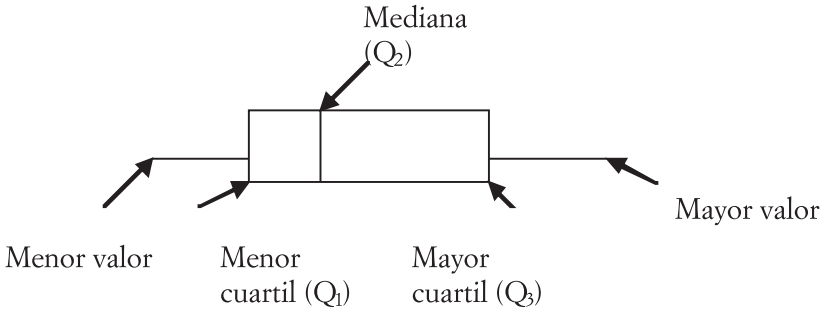


FIGURA 3-8

Gráfico de tallo y hojas

El gráfico de tallo y hojas se utiliza para variables cuantitativas, tanto continuas como discretas. Esta representación se basa en la ordenación de los datos a manera de gráfico, pero sin llegar a ello, utilizando las decenas y las unidades. Representamos los datos separando las decenas de las uni-

dades; por ejemplo, 63 se representa por 6 | 3. Las decenas se ponen en una columna, en forma vertical y las unidades a su derecha.

Por ejemplo, con los siguientes datos, resultaría el siguiente diagrama de tallo y hojas:

51, 64, 65, 54, 70 , 100, 25, 28 , 34, 56, 52, 73, 44, 48, 31, 67, 61, 60, 75, 72, 81, 82 ,94, 95, 68

2	5
3	4 1
4	4 8
5	1 4 6 2
6	4 5 7 1 0 8
7	0 3 5 2
8	1 2
9	4 5
10	0

El renglón 3| 4 1 quiere decir que en la lista de datos figura el 34 y el 31.

Los tallos se pueden representar también en intervalos; por ejemplo

TABLA 3-1: *Ejemplo de tallo y hojas*

Tallo	Hojas
0	6 7
1	2 9 0 5 2
2	3 5 1

En la tabla 3-1 :

- tallo 0 representa el intervalo de clase de 0 a 9;
- tallo l representa el intervalo de clase de 10 a 19; y
- tallo 2 representa el intervalo de clase 20 a 29.

Si al agrupar en intervalos las hojas están demasiado llenas, se pueden dividir los intervalos en dos o más componentes, para clarificar la información. Por ejemplo un intervalo 0-9 puede dividirse en dos intervalos de 0-4 y 5-9.

En la representación de *tallo y hojas* cada renglón es una posición de tallo y cada dígito de la derecha es una hoja. El procedimiento para realizarla es primero empezar con los tallos, es decir la columna de la izquierda, y después dato por dato ir llenando las hojas a la derecha de la línea vertical, en el tallo correspondiente. El tallo recoge todos los dígitos del número, menos el último que es la hoja; es decir, la hoja siempre tiene un solo dígito. Los números con decimales se redondean hasta el entero más cercano. El gráfico muestra como los datos se distribuyen (el mayor, el menor, el mas frecuente y los datos extremos). La mayor ventaja es que se resumen los datos al mismo tiempo que permanecen todos a la vista. La visualización de los datos se puede presentar girando el gráfico, con lo que resulta más significativa.

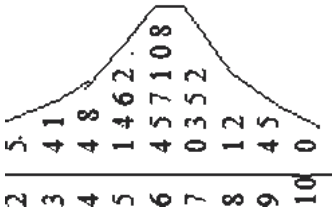


FIGURA 3-9

3.4. CORRELACIONES

Hasta ahora se han estudiado las características de una variable, pero en el análisis que estemos realizando puede ser necesario comprobar si existe alguna relación entre dos variables o varias variables. En este caso se realizan análisis bivariantes (dos variables) o multivariantes (varias variables). En el caso de que haya relación habrá que averiguar en qué grado se produce y cómo es.

Coeficiente de correlación lineal

Es una medida del grado de asociación lineal entre las variables X e Y. Se representa por $r = \frac{S_{xy}}{S_x \cdot S_y}$,

donde s_x, s_y son las desviaciones típicas de las variables X e Y respectivamente, y S_{xy} es la covarianza muestral de X e Y, que se define como la media de los productos de las desviaciones correspondientes de X e Y y de sus medias muestrales. $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Los valores de r se encuentran siempre entre –1 y 1. Cuando es cercano a 0 hay muy poca relación lineal, pero puede haberla de otro tipo. Si tiene un valor cercano a 1 o –1, la relación lineal es alta. Pero aunque el coeficiente indique correlación hay que buscar un sentido y explicación a la misma.

Matriz de correlaciones

Sea una muestra con n variables (X1, X2, ... Xn). Podemos ordenar en una matriz los diferentes coeficientes de correlación de cada variable con el resto y consigo misma, obteniendo una matriz con cada elemento igual a: $r_{ij} = \frac{S_{xy_{ij}}}{s_{x_i} \cdot s_{y_j}}$. Esta matriz se caracteriza por ser una matriz simétrica con la diagonal igual a 1.

Diagrama de dispersión

El diagrama de dispersión se conoce también como nube de puntos. Se utiliza para representar dos o más variables relacionadas sobre unos ejes cartesianos. En el eje de abscisas representamos los valores de X y en el de ordenadas los valores de Y, de tal forma que cada par viene representado por un punto del plano (X,Y). Normalmente se coloca la variable dependiente en el eje Y, mientras que en el X se coloca la variable independiente.

En el caso de que las dos variables estén agrupadas en intervalos, el diagrama se construye mediante casillas que tienen dentro tantos puntos como el valor de la frecuencia absoluta correspondiente a los intervalos X e Y.

Si las variables que componen el par son una discreta y otra continua, se utilizan las marcas de clase, siendo un caso similar al primero.

La información que proporciona el diagrama de dispersión revela aspectos importantes de la relación entre las variables, que pueden ser muy

útiles para enfocar o dirigir la investigación sobre la muestra, tales como: correlación entre los datos, relaciones positivas o directas, o bien relaciones negativas o inversas entre las variables, dispersión, tendencias no lineales, extensión de los datos o datos extremos.

Cuando los datos forman una línea, es clara la *relación lineal* entre las variables, es fuerte y hay alta correlación.

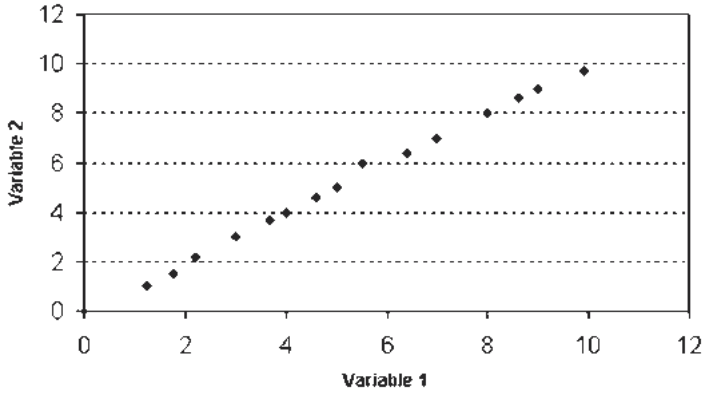


FIGURA 3-10

Si los puntos se agrupan en torno a una línea que va desde izquierda abajo hasta la derecha arriba, entonces se considera que la relación es positiva o directa, cuando mas cerca están los puntos de la línea mayor es la relación.

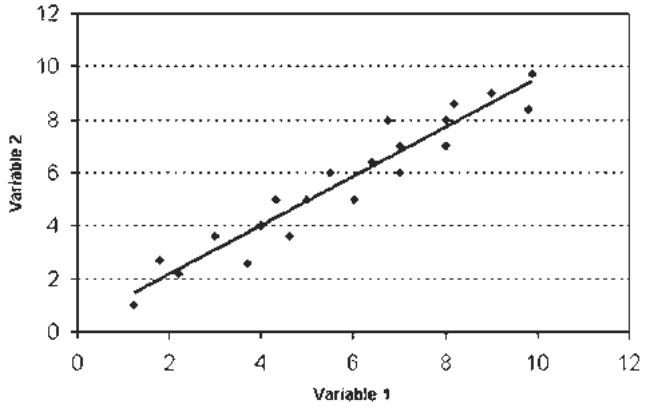


FIGURA 3-11

Si por el contrario, la línea en torno a la que se agrupan los puntos va desde arriba a la izquierda hasta abajo a la derecha, la relación entre las variables es negativa o inversa.

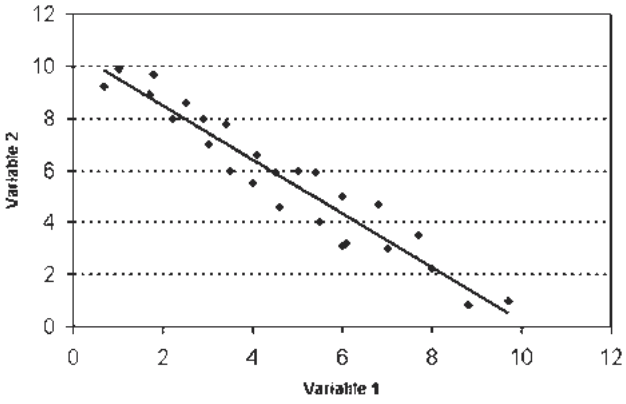


FIGURA 3-12

Si los puntos están aleatoriamente distribuidos, no hay relación entre las dos variables, esto quiere decir que no hay correlación entre ellas o que es muy pequeña.

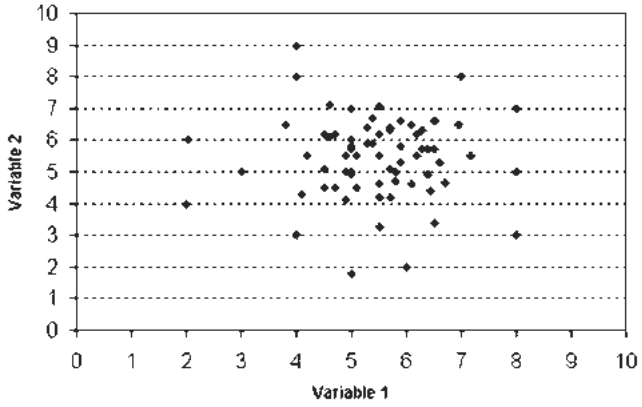


FIGURA 3-13

La relación entre las variables puede no ser lineal, sino que los puntos se agrupan, por ejemplo, en torno a una curva parábola, etc.

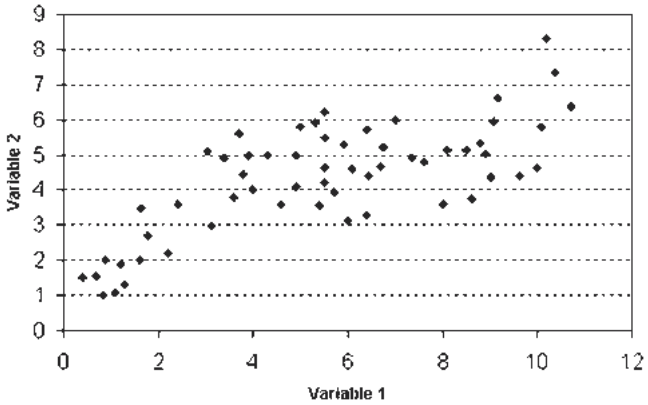


FIGURA 3-14

Un gráfico de dispersión también muestra si los datos están ampliamente dispersos (Fig 3-15) o concentrados (Fig.3-16), o si hay algún dato extremo que difiere de la pauta que presenta el resto (Fig. 3-17).

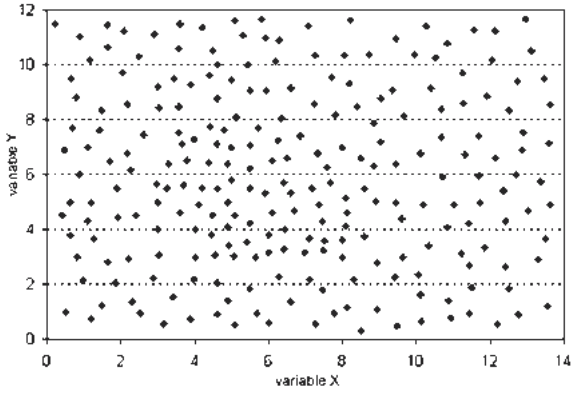


FIGURA 3-15

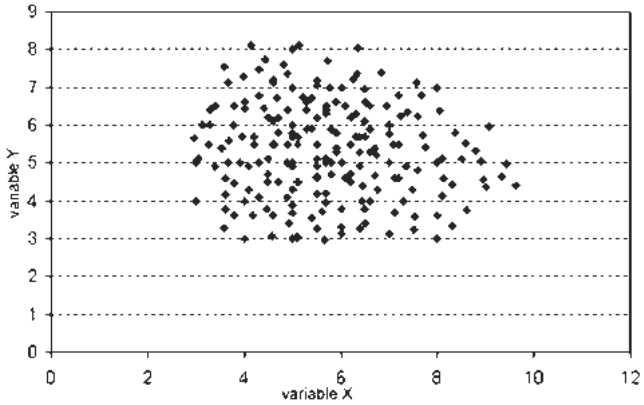


FIGURA 3-16

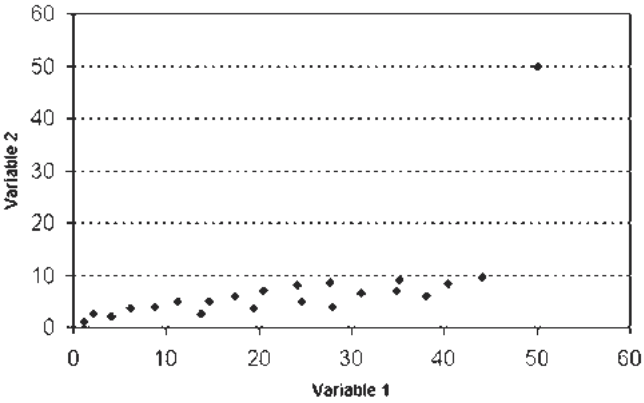


FIGURA 3-17

3.5. PRUEBAS DE NORMALIDAD

La distribución normal, también llamada gaussiana o de Gauss, es una de las que aparece con mas frecuencia en estadística. Hay muchas variables naturales asociadas a esta forma de distribución (talla, peso, coeficiente intelectual, etc.); precisamente por ello se la denomina *normal,* como ya hemos indicado con anterioridad. Muchos de los procesos estadísticos exigen que los datos sigan una distribución normal, por ejemplo en el análisis de covarianzas, por ello es importante comprobar que efectivamente es así. En muchas ocasiones la simple visión de los datos, como el histograma o la distribución de frecuencias, nos dice si la distribución es normal. Pero en ocasiones con esto no es suficiente y

son necesarios otros contrastes de hipótesis para comprobar si la muestra sigue una distribución normal. Cuando no sea así se podrán transformar o emplear otros métodos que no exijan normalidad en la distribución. No existe una única distribución normal, sino todo un conjunto de distribuciones con una forma común (la forma acampanada), diferenciadas por los valores de su media y su varianza. De entre todas ellas, la más utilizada es la **distribución normal estándar**, que corresponde a una distribución de media 0 y varianza 1 N(0,1).

Entre las herramientas para averiguar la normalidad de una distribución se encuentran:

- Los **gráficos de probabilidad normal** constituyen una de las alternativas. Este sistema consiste en comparar los datos de la muestra con los datos teóricos que corresponderían a una distribución normal, todo ello en un mismo gráfico. Si la distribución de la variable que estudiamos coincide con la normal, los puntos se concentrarán en torno a una línea recta, no obstante hay que tener presente que puede observarse mayor variabilidad en los extremos. Esto se puede hacer comparando las proporciones acumuladas de una variable con las de una distribución normal; o bien, comparando los cuantiles de la distribución que se analiza respecto de los de la normal. Con este sistema se puede también conocer porqué hay desviaciones respecto a la normal, o donde se producen. Una curva en forma de "U" o con alguna curvatura, significa que la distribución es asimétrica con respecto a la normal; mientras que un gráfico en forma de "S" significará que la distribución tiene colas mayores o menores que la normal; esto es, que existen pocas o demasiadas observaciones en las colas de la distribución.

- El **Q-Q plot** permite comparar la función de repartición de la muestra (en abscisa) con aquella que tendría una ley normal de media igual y varianza igual (en ordenadas). En el caso de una muestra procedente de una distribución normal, se debe observar una alineación casi perfecta con la primera bisectriz del plano. En el caso contrario, se deben observar desviaciones.

- Los coeficientes de asimetría y curtosis, antes expuestos, ayudan en este estudio.

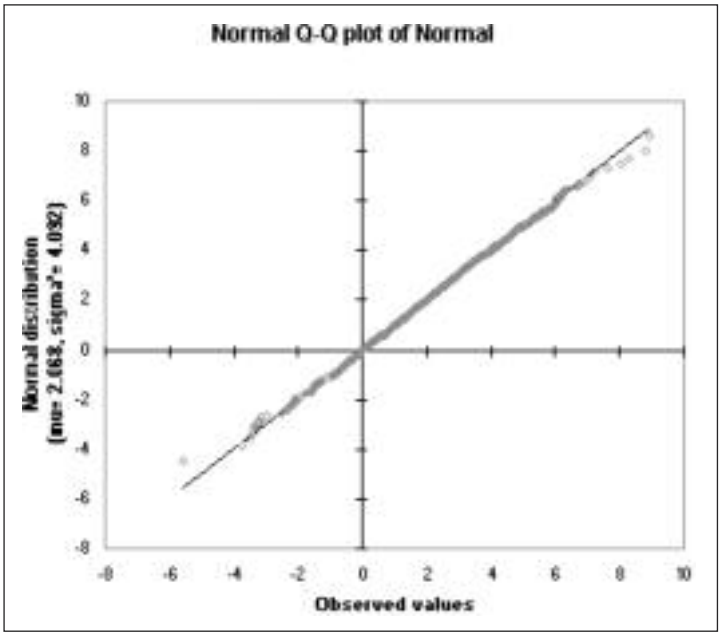


FIGURA 3-18

- **Contraste χ^2 de Pearson**. Para realizar este test se agrupan los datos en *k* clases (*k≥5*), como si fuéramos a construir un histograma, cubriendo todo el rango posible de valores, siendo deseable disponer, aproximadamente, del mismo número de datos en cada clase y al menos de tres datos en cada una. Llamamos *Oi* al número de datos observado en la clase *i*. Mediante el modelo de probabilidad que se desea verificar se calcula la probabilidad *Pi* asignada a cada clase, y por lo tanto, para una muestra de *n* datos, la frecuencia esperada según ese modelo de probabilidad es *Ei=n.Pi*. Se calcula entonces el siguiente índice de discrepancia entre las frecuencias observadas y las que era previsible encontrar si el modelo fuera el adecuado:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

que se distribuye aproximadamente como una χ^2 si el modelo es correcto.

Si el modelo se especifica de forma completa con las probabilidades *Pi*, conocidas antes de tomar los datos, el número de grados de libertad es *k-1*. Pero si se han estimado *r* parámetros del modelo a partir de los datos, entonces los grados de libertad son *k-r-1*.

- El **test de Kolmogorov-Smirnov.** Este contraste, que es válido únicamente para variables continuas, se basa en medir la máxima diferencia entre la distribución acumulada de los datos observados con la de una distribución normal, representada habitualmente como *D*. Si esta diferencia es significativa, se deberá rechazar la hipótesis de que la distribución es normal. Se compara el resultado con unos valores que se obtienen de una tabla de probabilidad desarrollada por Massey en 1951. Se proporciona un valor de probabilidad *P*, que corresponde, si estamos verificando un ajuste a la distribución normal, a la probabilidad de obtener una distribución que discrepe tanto como la observada si verdaderamente se hubiera obtenido una muestra aleatoria, de tamaño *n*, de una distribución normal. Si esa probabilidad es grande no habrá razones estadísticas para suponer que nuestros datos no proceden de una distribución normal, mientras que si es muy pequeña, no será aceptable suponer ese modelo probabilístico para los datos. Estos valores de probabilidad son válidos cuando la media y la distribución normal se conocen a priori. Cuando no es así son necesarias nuevas condiciones más complejas. El test de Kolmogorov-Smirnov otorga un peso menor a las observaciones extremas y por la tanto es menos sensible a las desviaciones que normalmente se producen en estos tramos que otros tests. No obstante hay otras pruebas mas precisas como la de Shapiro-Wilks o Jarque-Bera.

- El **test de Shapiro-Wilks** es uno de los más utilizados y se incorpora en el análisis de la mayoría de los programas estadísticos. En escala probabilística normal se representa en el eje horizontal, para cada valor observado en nuestros datos, la función de distribución o probabilidad acumulada observada, y en el eje vertical la prevista por el modelo de distribución normal. Si el ajuste es bueno, los puntos se deben distribuir aproximadamente según una recta a 45º.

En cualquier caso siempre es adecuado efectuar una representación gráfica de tipo histograma de los datos, y comparar el valor de la media y la mediana, así como evaluar el coeficiente de asimetría y apuntamiento, además de llevar a cabo una representación en escala probabilística de la distribución de probabilidad esperada versus observada, como la de la figura.

El proceso de cálculo es el siguiente:

Dada una muestra aleatoria simple de tamaño *n*,

$$(x_1, x_2, \ldots, x_n),$$

1. Se ordena la muestra de menor a mayor, obteniendo el nuevo vector muestral

$$(x_{(1)}, x_{(2)}, \ldots, x_{(n)}),$$

siendo $x_{(i)}$ el i-ésimo valor muestral tras la ordenación.

2. Se calcula el estadístico de contraste

$$W = \frac{1}{(n-1)s^2}\left[\sum_{i=1}^{k} a_{i,n}\left(x_{(n-i+1)} - x_{(i)}\right)\right]^2$$

siendo s^2 la varianza muestral,

$$k = \begin{cases} \frac{n}{2}, & \text{si } n \text{ es par} \\ \frac{n-1}{2}, & \text{si } n \text{ es impar} \end{cases}$$

y las $a_{i\,n}$ están tabuladas.

3. La distribución del estadístico W se encuentra también tabulada para cada nivel de significación.

El contraste de normalidad se plantea en los siguientes términos:

H_0: «la muestra procede de una población normal»

frente a la alternativa:

H_1: «la muestra no procede de una población normal».

- **Prueba de Jarque-Bera.** Con la prueba de Shapiro-Wilk, el riesgo de equivocarse rechazando la hipótesis es más importante que con la prueba de Jarque-Bera, especialmente cuando el tamaño de la muestra es muy grande. La prueba de Shapiro-Wilk funciona bien en muestras pequeñas, para valores de n menores a 30, y tiene serios problemas en muestras mayores a 5000 individuos. El cálculo de la prueba de normalidad de Jarque-Bera sigue los siguientes pasos:

$$JB = n\left[\frac{\hat{S}_K^2}{6} + \frac{(\hat{K}-3)^2}{24}\right]$$; para muestras grandes sigue una distribución

$$\chi^2_2$$

Siendo:

$$S_x = \frac{\left[E(X-\mu)^3\right]^2}{\left[E(X-\mu)^2\right]^3}$$

$$K = \frac{E(X-\mu)^4}{\left[E(X-\mu)^2\right]^2}$$

Para una variable aleatoria distribuida normalmente:

$$E(X-\mu)^3 = 0$$
$$E(X-\mu)^4 = 3\sigma^4$$

Por lo que en condiciones de normalidad:

$$S_x = 0$$
y
$$K = 3$$

Y en función de la muestra:

$$\hat{S}_x = \frac{\left[\sum\left(X_i - \bar{X}\right)^3 / (n-1)\right]^2}{\left[\sum\left(X_i - \bar{X}\right)^2 / (n-1)\right]^3}$$

$$\hat{K} = \frac{\sum \left(X_i - \bar{X} \right)^4 / (n-1)}{\left[\sum \left(X_i - \bar{X} \right)^2 / (n-1) \right]^2}$$

Posibles soluciones cuando se rechaza la hipótesis de normalidad

En el caso de que los análisis anteriores resulten en un rechazo de la normalidad en los datos de la muestra, existen soluciones:

- Si la distribución es más apuntada que la normal (mayor parte de los valores agrupados en torno de la media y colas más largas en los extremos), se debe investigar la presencia de heterogeneidad en los datos y de posibles valores atípicos o errores en los datos. La solución puede ser emplear pruebas no paramétricas.

- Si la distribución es unimodal y asimétrica, la solución más simple y efectiva suele ser utilizar una transformación para convertir los datos en normales.

- Cuando la distribución no es unimodal hay que investigar la presencia de heterogeneidad, ya que en estos casos la utilización de transformaciones no es adecuada y los métodos no paramétricos pueden también no serlo.

Uno de los sistemas mas frecuentes para conseguir la trasformación de los datos de forma que se ajusten a una normal es trasformarlos en logaritmos. En el caso de que algunos valores sean cero o muy pequeños, tendríamos problemas con la trasformación logarítmica, por lo que en lugar del ln(x) se puede emplear el ln(x+1). Cuando los datos provienen de recuentos, será más conveniente utilizar $\sqrt{x}$. Otra transformación habitualmente empleada es 1/x, que también precisa que sumemos una cantidad a cada valor si existen ceros.

Estas tres transformaciones comprimen los valores altos de los datos y expanden los bajos, en sentido creciente en el siguiente orden: $\sqrt{x}$ (la que menos), ln x, 1/x.

Si la concentración de datos está en el lado de la derecha de la distribución y la cola en la izquierda, se puede utilizar la transformación x^2, que comprime la escala para valores pequeños y la expande para valores altos.

Cuando los datos son proporciones o porcentajes de una distribución binomial, las diferencias con una distribución normal son más acusadas para valores pequeños o grandes de las proporciones, utilizándose entonces transformaciones basadas en $\arcsen \sqrt{p}$.

En todos los casos para los cálculos estadísticos basados en la teoría normal que presenten problemas, se utilizarán los valores transformados, pero después para la presentación de los resultados se efectuará la transformación inversa para presentarlos en su escala de medida natural.

En otros casos puede ser preferible utilizar sobre la muestra pruebas no paramétricas que no exigen la normalidad de los datos. También pueden servir para corroborar los resultados obtenidos a partir de la utilización de la teoría basada en la normal.

3.6. FUNCIONES Y HERRAMIENTAS DE EXCEL Y XLSTAT PARA LA EXPLORACIÓN DE DATOS

Excel y XLSTAT disponen de dos herramientas que sintetizan los análisis de exploración de datos que hemos expuesto en las líneas anteriores. En los ejemplos que se presentan en las líneas siguientes se explica el funcionamiento de estas herramientas, utilizando datos que tienen que ver con las políticas de familia.

EJERCICIO 3-1. Una de las variables de especial importancia en una de las políticas de conciliación de la vida familiar y laboral es la tasa de empleo de las mujeres en edad fértil y el número de hijos. Tenemos los siguientes datos de tasa de empleo para las mujeres europeas entre 20 y 49 años con 3 o mas hijos: Hungría 12.6; La República Checa 22; Eslovaquia 27.4; Francia 39.8; Reino-Unido 37.9; Alemania 37.9; Italia 35; Bélgica 49.2; Polonia 44.7; Chipre 51.6; Grecia 39.6; Austria 57.4; Finlandia 56.2; España 41.3; Portugal 60.2; Los Países Bajos 58.5; Dinamarca 67.2.

Con los datos en la hoja de Excel abrimos XLSTAT en descripción de datos y ponemos el cursor sobre la pestaña de estadísticas descriptivas, tal y como aparece en la figura adjunta.

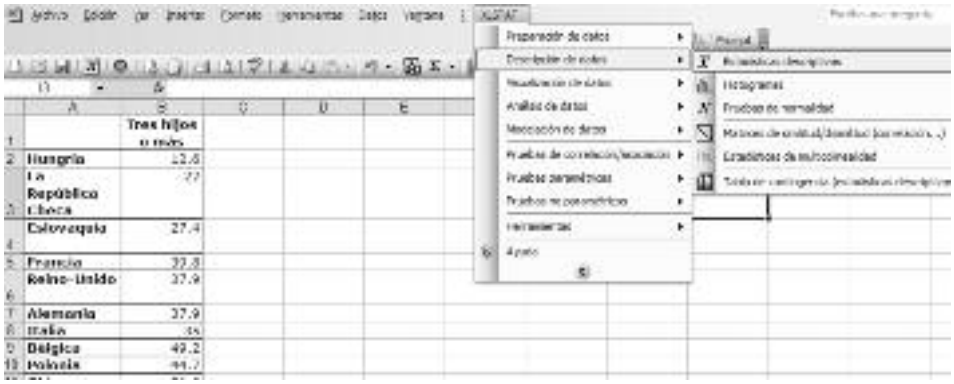


FIGURA 3-19

Al pinchar en estadísticas descriptivas se despliega la caja que aparece en la figura adjunta (Fig. 3-20). Se trata de datos cuantitativos, por lo que se marca esta opción, se pincha en el casillero situado justo debajo y se despliega el cursor (ratón) por los datos para que aparezcan reflejados en el casillero correspondiente.

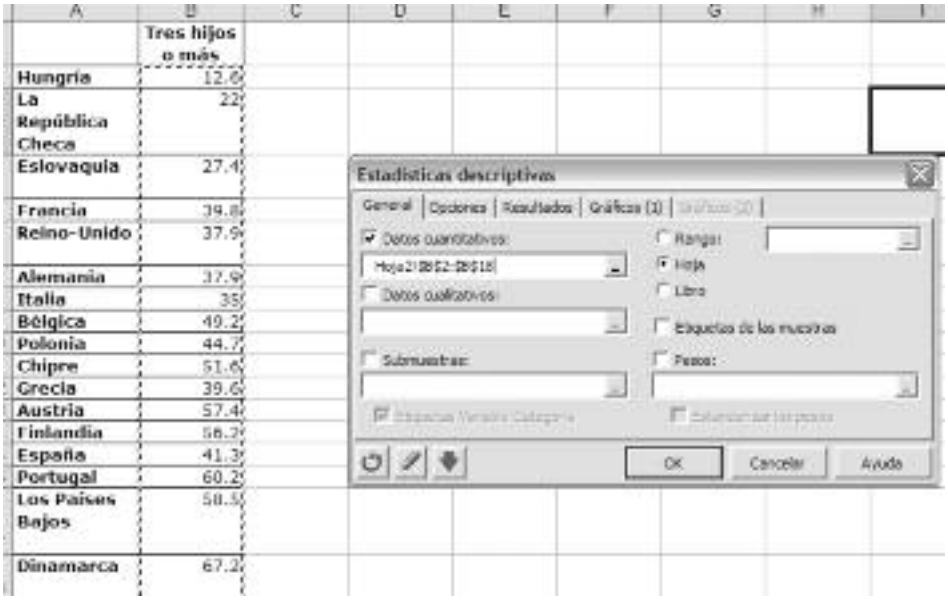


FIGURA 3-20

Si se pincha en la pestaña de *opciones*, se obtiene la imagen que aparece en la figura siguiente (Fig 3-21). Ahí marcamos la opción de estadísticas descriptivas y gráficos, así obtenemos parte de los gráficos que nos hacen falta en nuestro análisis.

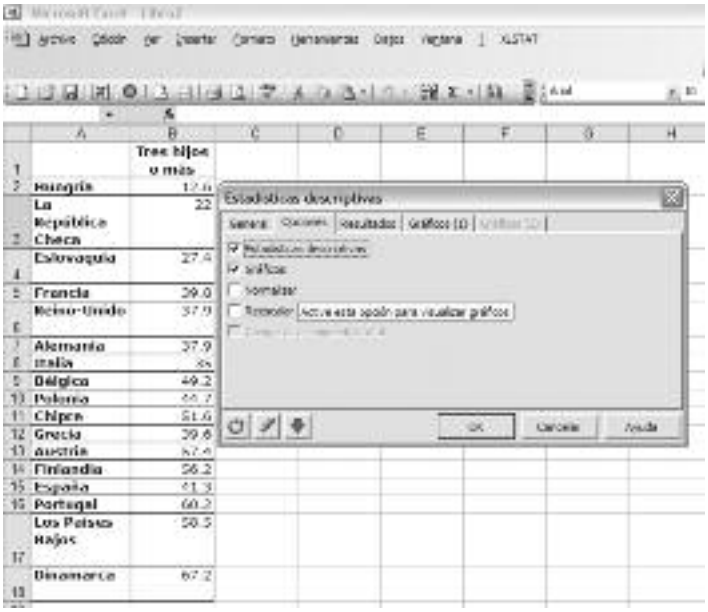


FIGURA 3-21

En la opción de resultados, escogemos «todas» en estadísticas descriptivas, según se indica en la figura adjunta (Fig 3-22).

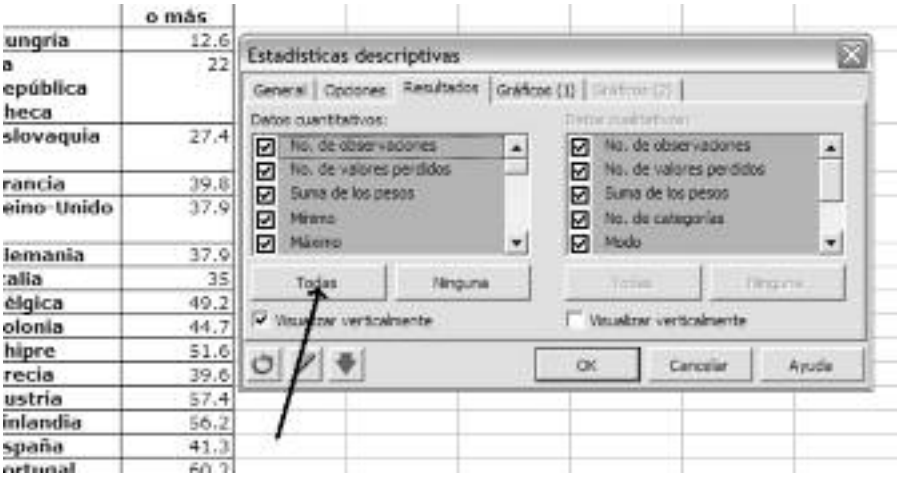


FIGURA 3-22

A continuación pinchamos en la pestaña de gráficos según aparece en la figura 3-23 adjunta. Los *scattergrams* (gráficos de dispersión) se utilizan para análisis bivariados, pero también pueden ser útiles en los análisis de una variable para dar una idea de la distribución y posible pluralidad de modas de una muestra. El *strip plot* representa los datos de la muestra como «tiras». Para un intervalo dado, cuanto más densas o más firmes sean las tiras, mas datos tendrán. Los *box plots* son los gráficos de caja y bigotes y los *Stem - and - leaf - plots* los gráficos de tallo y hojas que hemos descrito en las páginas anteriores. Los *gráficos P-P* (Probabilidad-Probabilidad) y *Q-Q* (Quantile-Quantile) son los utilizados para comparar la distribución de la muestra con la distribución normal, con la misma media y desviación. Si la muestra se distribuye como una normal, los datos caerán a lo largo del primer bisector del plano.

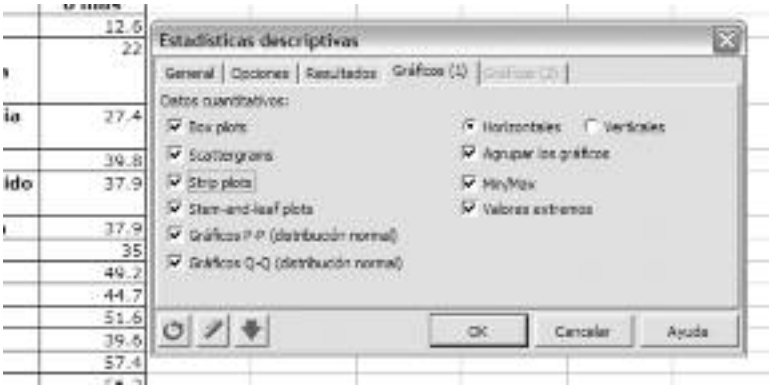


FIGURA 3-23

Ya solo queda pulsar en OK, tras lo cual aparece un cuadro en el que se presenta un resumen de los datos para verificar que es correcta la selección inicial realizada.

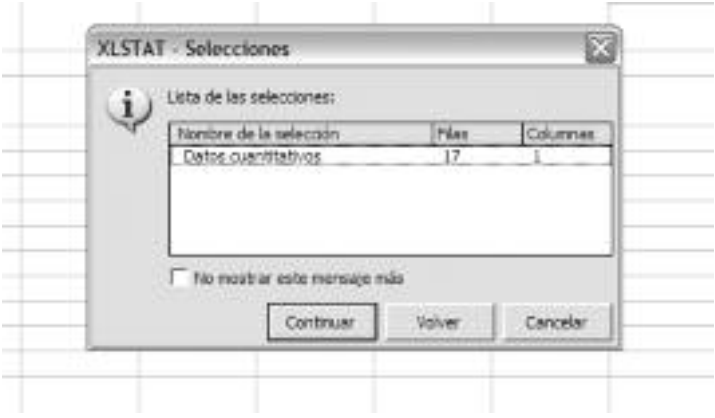


FIGURA 3-24

Y en una nueva hoja del libro de Excel, con la denominación «Desc» aparecen los resultados, que podemos copiarlos en Word simplemente seleccionándolos en Excel, copiando y pegando en Word:

FIGURA 3-2. *Estadísticas descriptivas ej. 3-1 (Datos cuantitativos)*

Estadística	*X1*
N.º de observaciones	17
N.º de valores perdidos	0
Suma de los pesos	17
Mínimo	12.600
Máximo	67.200
Frecuencia del mínimo	1
Frecuencia del máximo	1
Amplitud	54.600
1.° Cuartil	37.900
Mediana	41.300
3.° Cuartil	56.200
Suma	738.500
Media	43.441
Varianza (n)	197.914
Varianza (n-1)	210.284
Desviación típica (n)	14.068
Desviación típica (n-1)	14.501
Coeficiente de variación	0.324
Asimetría (Pearson)	–0.366
Asimetría (Fisher)	–0.402
Asimetría (Bowley)	0.628
Curtosis (Pearson)	–0.434
Curtosis (Fisher)	–0.138
Error estándar de la media	3.517
Límite inferior de la media (95%)	35.985
Límite superior de la media (95%)	50.897
Desviación absoluta media	11.467
Desviación absoluta mediana	10.300
Media geométrica	40.535
Desviación típica geométrica	1.520
Media armónico	36.707

Los resultados indican que como mínimo, las mujeres en Europa tienen una tasa de empleo, cuando tienen familias numerosas, de 12,6 por ciento (valor de Hungría) y como máximo de 67,2 por ciento (valor de Dinamarca). Entre ambos hay una amplitud importante de 54,6 puntos porcentuales. El primer cuartil se corresponde con el 37,9 por ciento, donde se sitúan Alemania y el Reino Unido, la Mediana o segundo cuartil en 41,3 que corresponde a España, y el tercer cuartil en 56,2, que corresponde a Finlandia.

La muestra parece presentar asimetría por la izquierda, y apuntamiento.

Y los gráficos:

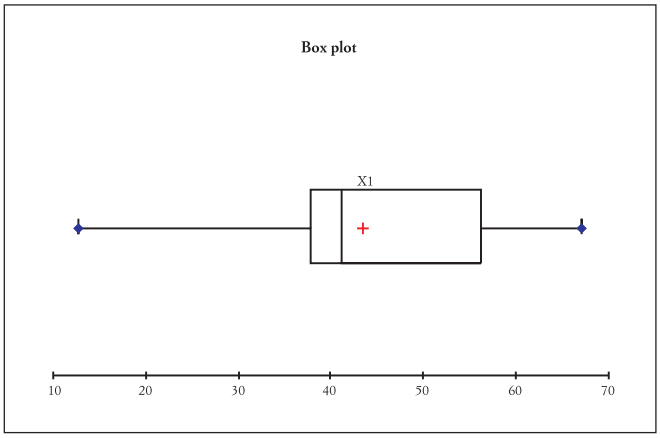


FIGURA 3-25

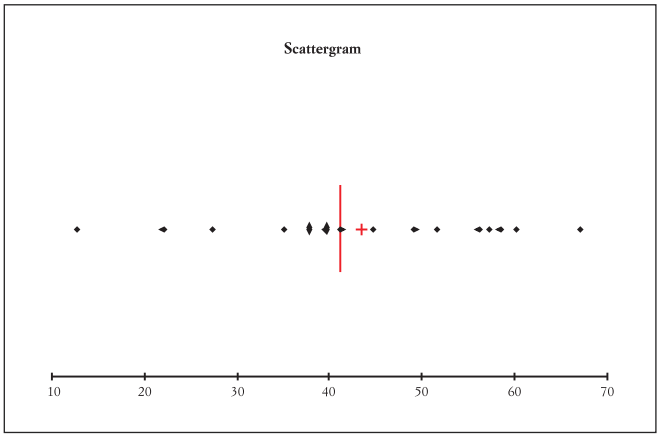


FIGURA 3-26

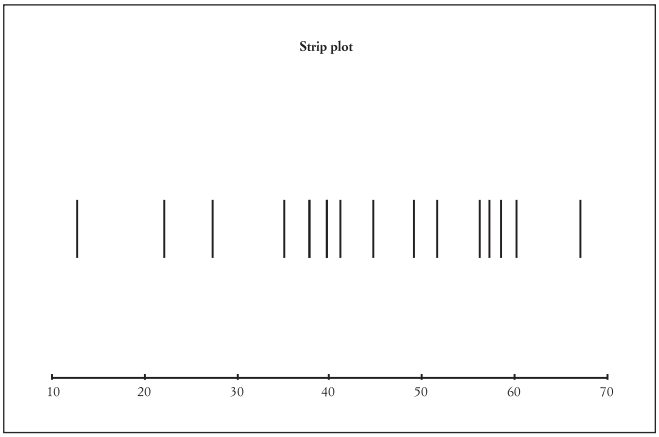


FIGURA 3-27

Stem -and-leaf plot (X1):	
Unidad: 10	
1	3
2	2 7
3	5 8 8 0 0
4	1 5 9
5	2 6 7 9
6	0 7

FIGURA 3-28

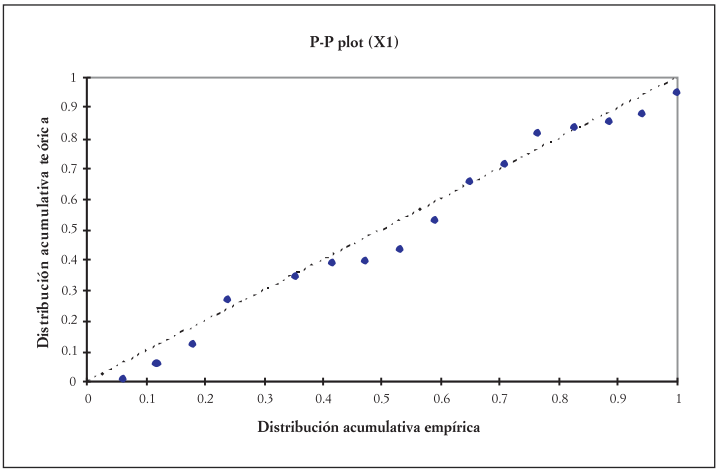


FIGURA 3-29

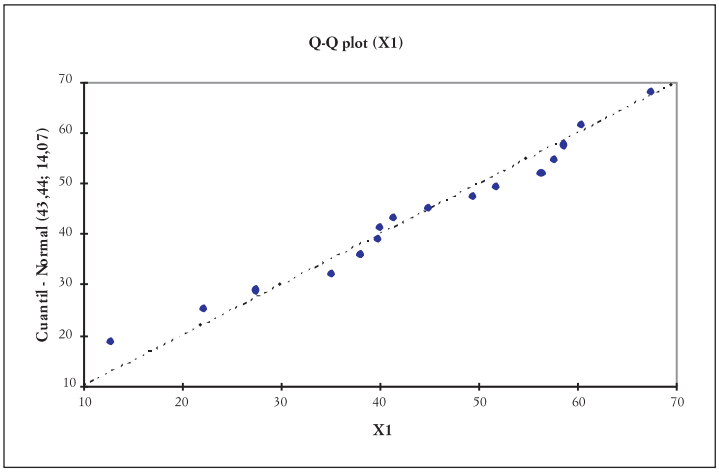


FIGURA 3-30

Los gráficos confirman la ligera concentración de datos a la izquierda de la mediana.

La información se complementaría con el histograma, que se obtiene al desplegar «descripción de datos», según como se recoge en la figura adjunta.

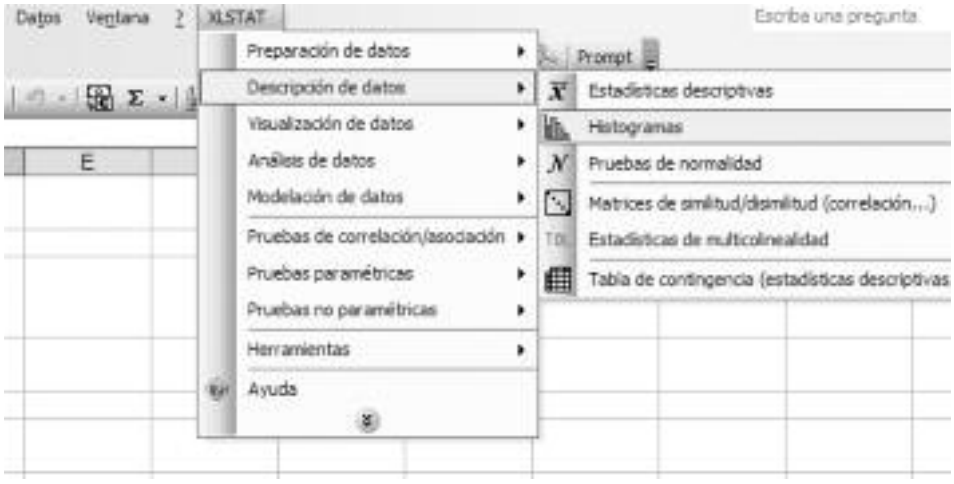


FIGURA 3-31

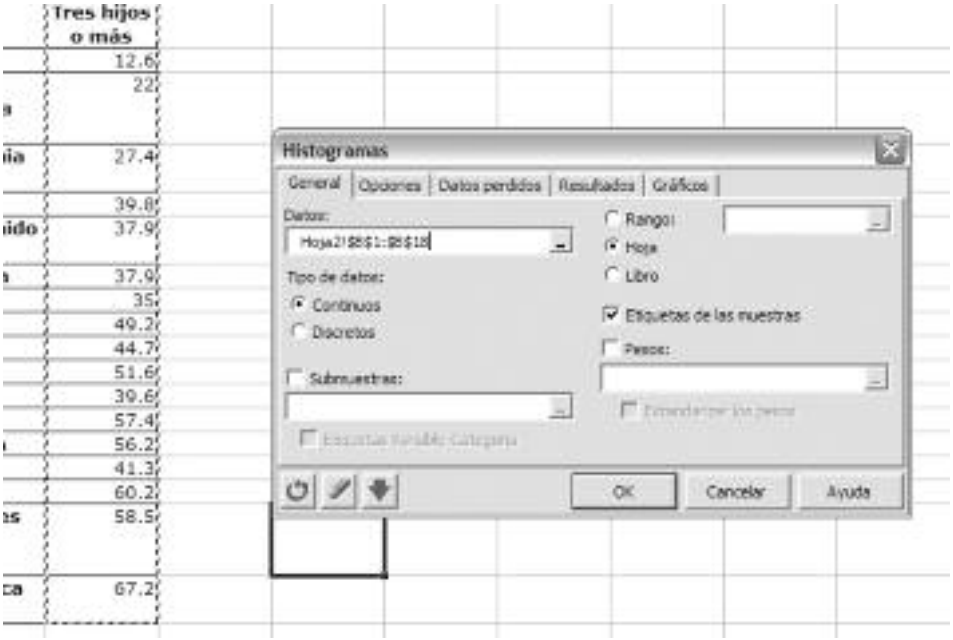


FIGURA 3-32

Pinchando en histogramas aparece el cuadro de la figura anterior, donde se seleccionan los datos.

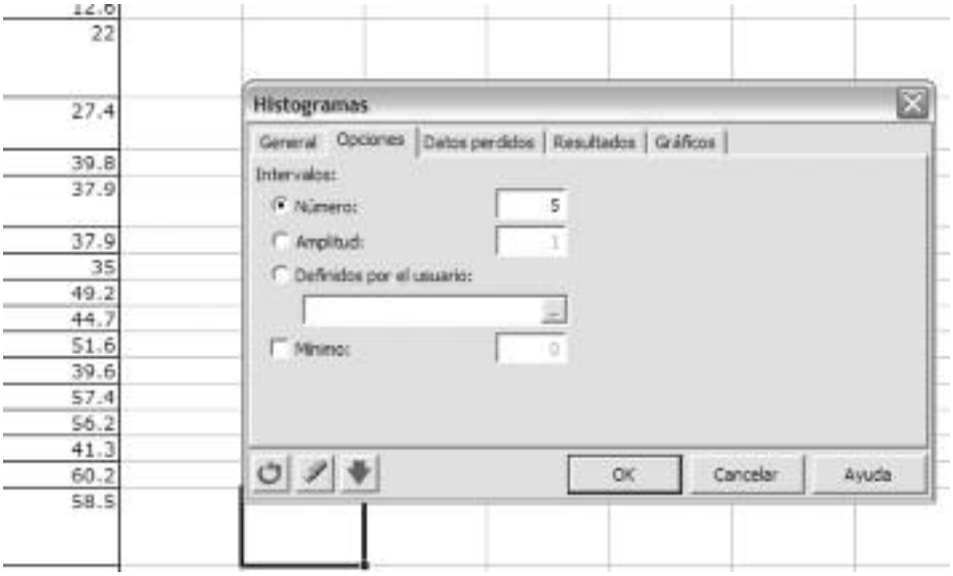


FIGURA 3-33

En intervalos se ha escogido la opción de 5, aplicando la regla de $1+3{,}3 \log_{10}$ (17). No se ha escogido la opción de que aparezcan en resultados las estadísticas descriptivas, puesto que ya las habíamos obtenido anteriormente. En gráficos escogemos histogramas de barras e histogramas acumulativos.

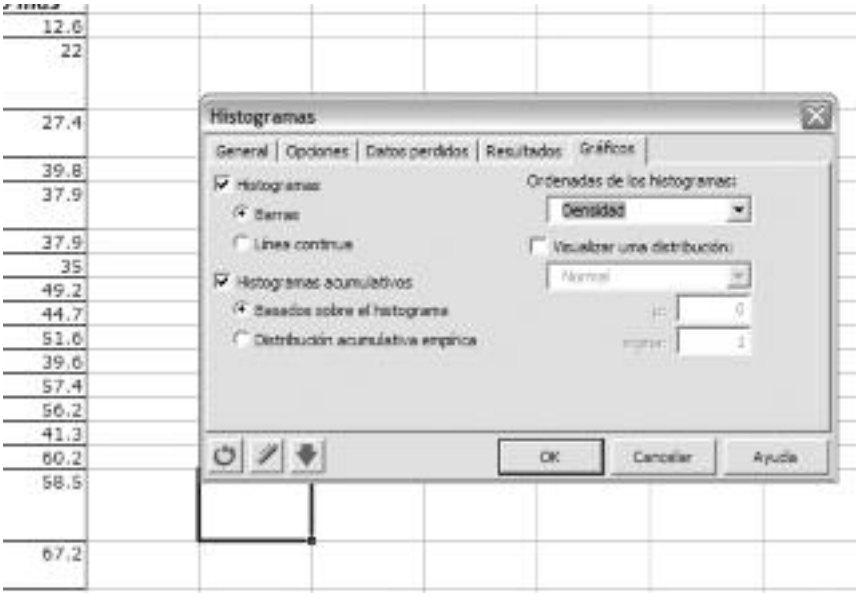


FIGURA 3-34

El resultado, que se origina en una nueva hoja del libro de Excel, con el nombre de histograma, es el que aparece en la figura siguiente:

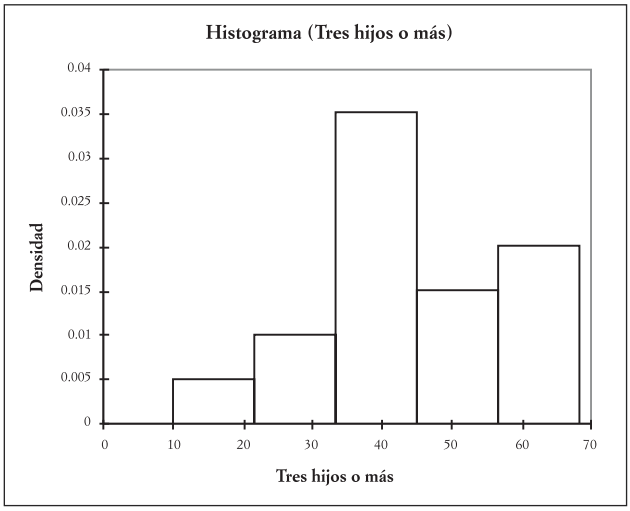


FIGURA 3-35

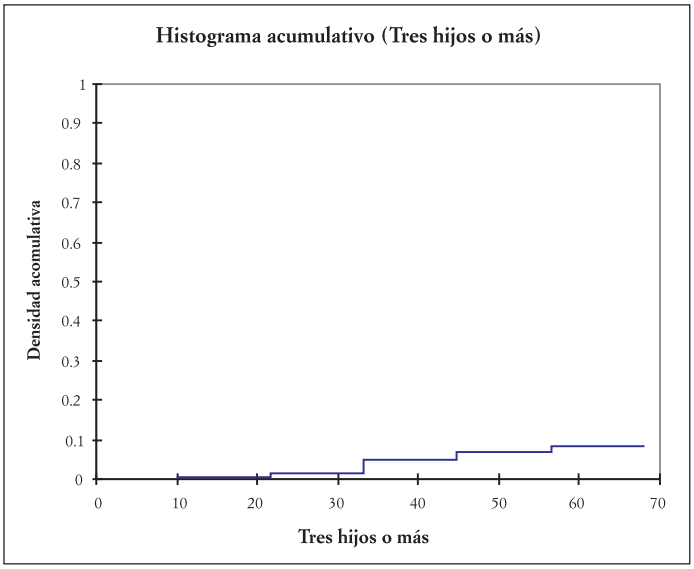


FIGURA 3-36

A fin de confirmar la normalidad de los datos, al desplegar descripción de datos, se pincha en *pruebas de normalidad*,

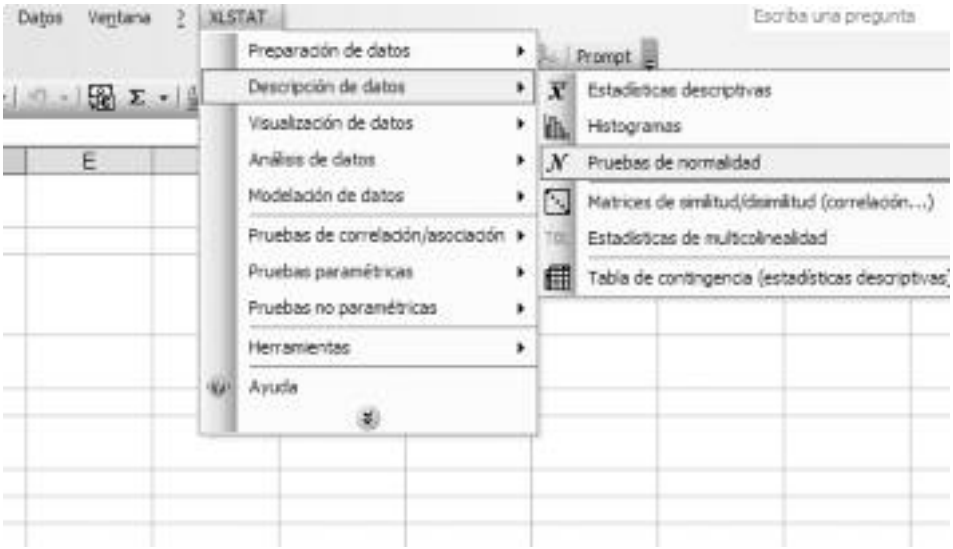


FIGURA 3-37

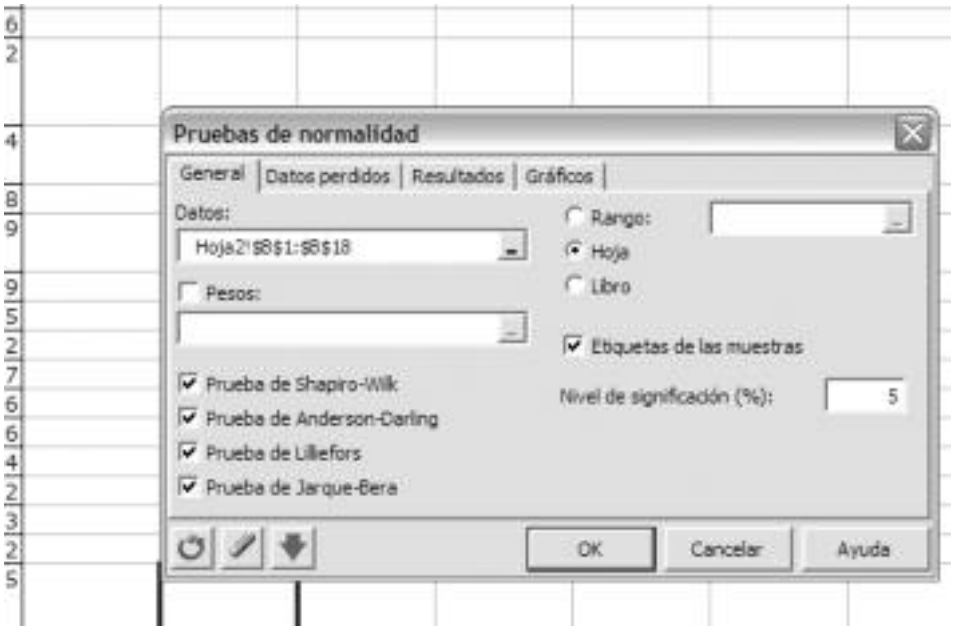


FIGURA 3-38

En la caja que aparece tras pinchar en pruebas de normalidad se señalan los datos que se incluyen en la prueba. Los gráficos ya están realizados en los análisis descriptivos anteriores, por ello no se vuelven a realizar. Se pincha en OK y el resultado que se obtiene es el que se recoge en las tablas siguientes:

TABLA 3-3. *Prueba de Shapiro-Wilk (Tres hijos o más)*

W	0.971
p-valor	0.836
alfa	0.05

Interpretación de la prueba de Shapiro-Wilk:

H0: La muestra sigue una ley Normal.

Ha: La muestra no sigue una ley Normal.

Como el p-valor calculado es 0,836, mayor que el nivel de significación alfa=0,05, se puede aceptar la hipótesis nula H0. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es de 83,59%.

TABLA 3-4. *Prueba de Anderson-Darling (Tres hijos o más)*

A≈	0.235
p-valor	0.754
alfa	0.05

Interpretación de la prueba de Anderson-Darling:

H0: La muestra sigue una ley Normal.

Ha: La muestra no sigue una ley Normal.

Como el p-valor calculado es 0,754, mayor que el nivel de significación alfa=0,05, se puede aceptar la hipótesis nula H0. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es de 75,41%.

TABLA 3-5. *Prueba de Lilliefors (Tres hijos o más)*

D	0.116
D (estandarizado)	0.478
p-valor	0.788
alfa	0.05

Interpretación de la prueba de Lilliefors:

H0: La muestra sigue una ley Normal.

Ha: La muestra no sigue una ley Normal.

Como el p-valor calculado 0,788 es mayor que el nivel de significación alfa=0,05, se puede aceptar la hipótesis nula H0. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es de 78,77%.

TABLA 3-6. *Prueba de Jarque-Bera (Tres hijos o más)*

JB (Valor observado)	0.690
JB (Valor crítico)	5.991
GDL	2
p-valor	0.708
alfa	0.05

Interpretación de la prueba de Jarque-Bera:

H0: La muestra sigue una ley Normal.

Ha: La muestra no sigue una ley Normal.

Como el p-valor calculado 0,708 es mayor que el nivel de significación alfa=0,05, se puede aceptar la hipótesis nula H0. El riesgo de rechazar la hipótesis nula H0 cuando es verdadera es de 70,81%.

Así pues, todas las pruebas realizadas conducen a confirmar que los datos siguen una distribución normal.

BIBLIOGRAFÍA

— Duncan Williamson, «Box and Whisker Diagrams». http://www.duncanwil.co.uk/boxplot.html

— Neville Hunt «Boxplots in Excel» http://www.mis.coventry.ac.uk/~nhunt/boxplot.htm Pérez C. (2002) Estadística Aplicada a través de Excel. Pearson. Prentice Hall. Madrid.

— Pérez C. (2005) Muestreo estadístico. Conceptos y problemas resueltos. Pearson. Prentice Hall.

CAPÍTULO IV
TRATAMIENTO DE VALORES PERDIDOS Y EXTREMOS

CÉSAR PÉREZ LÓPEZ
DANIEL SANTÍN GONZÁLEZ

4.1. TRATAMIENTO DE LOS VALORES PERDIDOS

El tratamiento de la información faltante constituye una de las tareas previas a cualquier análisis de investigación social basado en información sobre individuos o familias. Cuando se realiza una descripción de la muestra, se calcula un indicador simple o se aplica un método de análisis multivariante sobre los datos disponibles puede ser que no exista información para determinadas observaciones y variables. Estamos entonces ante valores ausentes, valores perdidos o valores *missing* término anglosajón con el que a menudo la literatura se refiere a este tipo de datos. La presencia de esta información faltante puede deberse a un registro defectuoso, a la ausencia natural de la información buscada o a una falta de respuesta (total o parcial). Dependiendo de la cantidad de datos perdidos y de su distribución en la base de datos en observaciones y columnas deberemos tomar una decisión sobre como reemplazar la información perdida.

Contrastes para valores perdidos

La primera prueba a realizar cuando existen datos *missing* es comprobar si se distribuyen aleatoriamente en todo el conjunto de datos. Es vital que el investigador averigüe si el proceso de ausencia de datos tiene lugar de forma aleatoria. Una primera prueba para valorar los datos ausentes para una única variable Y consiste en formar dos grupos de valores para Y, los que tienen datos ausentes y los que no los tienen. A continuación, para cada variable X distinta de Y, se realiza un test de diferencia de medias para determinar si existen diferencias significativas entre lo dos grupos de valores determinados por la variable Y (ausentes y no ausentes) sobre X. Si vamos considerando como Y cada una de las variables del análisis y repitiendo el proceso anterior y se encuentra que un porcentaje alto de las diferencias son no significativas, se puede concluir que los datos ausentes obedecen a un *proceso completamente aleatorio* y por tanto pueden realizarse análisis estadísticos fiables con nuestras variables *imputando los datos ausentes* por los métodos que se verán más adelante sin tener que eliminar estos datos.

También es habitual comprobar la distribución aleatoria de los datos *missing* mediante la *prueba de las correlaciones dicotomizadas*. Para realizar esta prueba, para cada variable *Y* del análisis se construye una variable dicotomizada asignando el valor cero a los valores ausentes y el valor uno a los valores presentes. A continuación se dicotomizan todas las variables del análisis y se halla su matriz de correlaciones acompañada de los contrastes de significatividad de cada coeficiente de correlación de la matriz. Las correlaciones indican el grado de asociación entre los valores perdidos sobre cada par de variables, con lo que se puede concluir que si los elementos de la matriz de correlaciones son en su gran mayoría no significativos, los datos ausentes son completamente aleatorios.

Adicionalmente existen pruebas formales de aleatoriedad de los datos *missing* como el *test conjunto de aleatoriedad de Little*, contraste formal basado en la Chi-cuadrado, cuyo *p*-valor indica si los valores perdidos constituyen o no un conjunto de números aleatorios.

A continuación se ilustran los conceptos anteriores con un ejemplo basado en los datos recogidos en un cuestionario con 6 preguntas sobre comportamientos y actitudes de 20 familias encuestadas. Las respuestas a las 6 preguntas se recogen en 6 variables (*V*1, *V*2, *V*3, *V*4, *V*5 y *V*6) cuyo rango varía entre 1 y 10 reflejando la valoración que el encuestado da a la característica que refleja la pregunta.

Los datos de las 6 variables en los 20 cuestionarios se recogen en la tabla de la Figura 4-1.

Cuestionario	V1	V2	V3	V4	V5	V6
1	5	6	2	1	.	5
2	7	.	4	5	5	7
3	.	1	5	8	5	8
4	3	5	1	.	7	5
5	5	5	8	3	7	8
6	5	1	.	1	2	8
7	4	.	2	8	9	8
8	5	1	9	1	1	9
9	7	5	1	1	1	.
10	2	2	1	4	6	6
11	9	1	1	.	7	5
12	5	5	8	9	9	5
13	.	9	1	9	7	9
14	5	6	2	1	1	5
15	7	7	4	5	4	7
16	1	1	5	8	5	.
17	3	5	1	7	.	5
18	5	5	.	3	7	8
19	5	1	1	1	2	8
20	5	1	9	1	1	9

FIGURA 4-1

Una vez tabulada la información, la primera tarea sería ver la tabla de frecuencias de los valores perdidos por variables para tener una idea de su magnitud. Para ello excel tiene la función CONTAR.BLANCO que utilizamos para saber el número de datos perdidos. A continuación se presenta dicha información (Figura 4-2), observándose que para todas las variables el porcentaje de valores perdidos es del 10%, 2 valores perdidos por columna, mientras que el de valores válidos es el 90%.

CONTAR.BLANCO =CONTAR.BLANCO(B2:B21)

	A	B	C	D	E	F	G
1	Cuestionario	V1	V2	V3	V4	V5	V6
2	1	5	6	2	1		5
3	2	7		4	5	5	7
4	3		1	5	8	5	8
5	4	3	5	1		7	5
6	5	5	5	8	3	7	8
7	6	5	1		1	2	8
8	7	4		2	8	9	8
9	8	5	1	9	1	1	9
10	9	7	5	1	1	1	
11	10	2	2	1	4	6	6
12	11	9	1	1		7	5
13	12	5	5	8	9	9	5
14	13		9	1	9	7	9
15	14	5	6	2	1	1	5
16	15	7	7	4	5	4	7
17	16	1	1	5	8	5	
18	17	3	5	1	7		5
19	18	5	5		3	7	8
20	19	5	1	1	1	2	8
21	20	5	1	9	1	1	9
22		=CONTAR.BLANCO(B2:B21)			2	2	2

FIGURA 4-2

El siguiente paso es determinar si los datos ausentes se distribuyen aleatoriamente. Para ello comparamos las observaciones con y sin datos ausentes para cada variable en función de las demás variables. La primera tarea será generar nuevas variables *V*11, *V*21, *V*31, *V*41, *V*51 y *V*61 (una para cada variable existente) asignándole el valor uno para datos válidos y el valor cero para datos ausentes. Este cálculo se puede realizar de diversas maneras aunque se recomienda utilizar la función SI(B2>0;1;0). Tendremos la tabla de la Figura 4-3.

Cuest.	V1	V2	V3	V4	V5	V6	V11	V21	V31	V41	V51	V61
1	5	6	2	1	.	5	1	1	1	1	0	1
2	7	.	4	5	5	7	1	0	1	1	1	1
3	.	1	5	8	5	8	0	1	1	1	1	1
4	3	5	1	.	7	5	1	1	1	0	1	1
5	5	5	8	3	7	8	1	1	1	1	1	1
6	5	1	.	1	2	8	1	1	0	1	1	1
7	4	.	2	8	9	8	1	0	1	1	1	1
8	5	1	9	1	1	9	1	1	1	1	1	1
9	7	5	1	1	1	.	1	1	1	1	1	0
10	2	2	1	4	6	6	1	1	1	1	1	1
11	9	1	1	.	7	5	1	1	1	0	1	1
12	5	5	8	9	9	5	1	1	1	1	1	1
13	.	9	1	9	7	9	0	1	1	1	1	1
14	5	6	2	1	1	5	1	1	1	1	1	1
15	7	7	4	5	4	7	1	1	1	1	1	1
16	1	1	5	8	5	.	1	1	1	1	1	0
17	3	5	1	7	.	5	1	1	1	1	0	1
18	5	5	.	3	7	8	1	1	0	1	1	1
19	5	1	1	1	2	8	1	1	1	1	1	1
20	5	1	9	1	1	9	1	1	1	1	1	1

FIGURA 4-3

Ahora consideramos los dos grupos formados en la variable *V*1 (valores válidos y valores ausentes) que vienen definidos por la variable *V*11 y hacemos un contraste de igualdad de medias (véase tema 10) para los dos grupos de valores definidos en cada una de las restantes variables (*V*2 a *V*6) por los valores de *V*11. Para ello utilizamos Herramientas → Análisis de Datos → Pruba t para dos muestras (suponiendo varianzas iguales o desiguales según el caso) o bien a través de la herramienta de XLSTAT → Pruebas paramétricas → Pruebas t y z para dos muestras.

Si el lector realiza el ejercicio se observará que para prácticamente todas las variables, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de cada una de ellas (los intervalos de confianza contienen el valor cero). Por lo tanto se puede concluir con bastante fiabilidad la distribución aleatoria de los datos perdidos, conclusión que permitirá realizar análisis estadísticos con los datos aplicando distintos métodos de imputación de la información faltante.

Para comprobar la aleatoriedad de los datos ausentes también se puede utilizar la matriz de correlaciones dicotomizadas. Se trata de calcular la matriz de correlaciones de las variables resultantes al sustituir lo valores perdidos de las variables iniciales por ceros, y los valores válidos por unos. En nuestro caso se trataría de hallar la matriz de correlaciones de las variables *V*11 a *V*61. Este análisis puede realizarse mediante Herramientas → Análisis de Datos → Coeficiente de Correlación o alternativamente con XLSTAT → Pruebas de correlación/asociación → Pruebas de correlación Tenemos los resultados de la Figura 4-4.

		V11	V21	V31	V41	V51	V61
V11	Correlación de Pearson	1	-,111	-,111	-,111	-,111	-,111
	p-valor	.	,641	,641	,641	,641	,641
V21	Correlación de Pearson	-,111	1	-,111	-,111	-,111	-,111
	p-valor	,641	.	,641	,641	,641	,641
V31	Correlación de Pearson	-,111	-,111	1	-,111	-,111	-,111
	p-valor	,641	,641	.	,641	,641	,641
V41	Correlación de Pearson	-,111	-,111	-,111	1	-,111	-,111
	p-valor	,641	,641	,641	.	,641	,641
V51	Correlación de Pearson	-,111	-,111	-,111	-,111	1	-,111
	p-valor	,641	,641	,641	,641	.	,641
V61	Correlación de Pearson	-,111	-,111	-,111	-,111	-,111	1
	p-valor	,641	,641	,641	,641	,641	.

FIGURA 4-4

Las correlaciones resultantes entre las variables dicotómicas indican la medida en que los datos ausentes están relacionados entre pares de variables. Las correlaciones bajas indican una baja asociación entre los procesos de ausencia de datos para esas dos variables. En nuestro caso todas las correlaciones son bajas y significativas, lo que corrobora la presencia de aleatoriedad de los datos ausentes.

Tratamiento de los valores perdidos: Imputación y supresión

Una vez que se ha contrastado la existencia de aleatoriedad en los datos ausentes ya se puede tomar una decisión para dichos datos antes de comenzar cualquier análisis estadístico o de investigación social con ellos.

Podemos comenzar incluyendo sólo en el análisis las observaciones (casos) con datos completos (filas cuyos valores para todas las variables sean válidos), es decir, cualquier fila que tenga algún dato desaparecido se elimina del conjunto de datos antes de realizar el análisis. Este método se suele denominar *aproximación de casos completos* o *supresión de casos según lista* y suele ser el método por defecto en la mayoría del *software* estadístico. Este método es apropiado cuando no hay demasiados valores perdidos, porque su supresión provocaría una muestra representativa de la información total. En caso contrario se reduciría mucho el tamaño de la muestra a considerar para el análisis y no sería representativa de la información completa. En el caso extremo podríamos imaginar que si siguiésemos esta técnica cuando tenemos un solo valor perdido en todas las filas que se repartiera uniformemente por todas las columnas ¡nos quedaríamos sin base de datos!.

Otro método consiste en la *supresión de datos según pareja*, es decir, se trabaja con todos los casos (filas) posibles que tengan valores válidos para cada par de variables que se consideren en el análisis independientemente de lo que ocurra en el resto de las variables. Este método elimina menos información y se utiliza siempre en cualquier análisis bivariante o transformable en bivariante.

Otro método adicional consiste en *suprimir los casos (filas) o variables (columnas)* que peor se comportan respecto a los datos ausentes. Nuevamente es necesario sopesar la cantidad de datos a eliminar. Debe siempre considerarse lo que se gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable o conjunto de casos en el análisis estadístico.

La alternativa a los métodos de supresión de datos es la *imputación de la información faltante*. La imputación es el proceso de estimación de valores ausentes basado en valores válidos de otras variables o casos de la muestra. A continuación se estudian diferentes métodos de imputación.

Un primer método de imputación no reemplaza los datos ausentes sino que imputa las características de la distribución (por ejemplo, la desviación típica) o las relaciones de todos los valores válidos disponibles (por ejemplo, correlaciones).

El proceso de imputación no consiste en reemplazar los datos ausentes por el resto de los casos, sino en utilizar las características de la distribución o las relaciones de todos los valores válidos posibles, como representantes para toda la muestra entera. Este método se denomina *enfoque de disponibilidad completa*.

Un segundo grupo de métodos de imputación ya son métodos de sustitución de datos ausentes por valores estimados sobre la base de otra información existente en la muestra. Consideraremos en este grupo el método de sustitución del caso, el método de sustitución por la media o la mediana, el método de sustitución por un valor constante, el método de imputación por interpolación lineal, el método de imputación por regresión y el método de imputación múltiple.

En el *método de imputación por sustitución del caso* las observaciones (casos) con datos ausentes se sustituyen con otras observaciones no muestrales. Por ejemplo, en una encuesta sobre hogares a veces se sustituye un hogar de la muestra que no contesta por otro hogar que no está en la muestra y que probablemente contestará. Este método de imputación suele utilizarse cuando existen casos con todas sus observaciones ausentes o con la mayoría de ellas.

En el *método de imputación de sustitución por la media* los datos ausentes se sustituyen por la media de todos los valores válidos de su variable correspondiente. Este método tiene la ventaja de que se implementa fácilmente y proporciona información completa para todos los casos, pero tiene la desventaja de que modifica las correlaciones e invalida las estimaciones de la varianza derivadas de las fórmulas estándar de la varianza para conocer la verdadera varianza de los datos.

Cuando hay valores extremos en las variables, se sustituyen los valores ausentes por la mediana (en vez de por la media), ya que la mediana es un estadístico resumen de los datos más robusto. De esta forma se tiene el *método de imputación de sustitución por la mediana*.

A veces, cuando hay demasiada variabilidad en los datos, suele sustituirse cada valor ausente por la media o mediana de un cierto número de observaciones adyacentes a él. En este tipo de imputación suele incluirse también el *método de imputación por interpolación* en el cual se sustituye cada valor ausente de una variable por el valor resultante de realizar una interpolación con los valores adyacentes.

En el *método de imputación de sustitución por valor constante* los datos ausentes se sustituyen por un valor constante apropiado derivado de fuentes externas o de una investigación previa. En este caso el investigador debe asegurarse de que la sustitución de los valores ausentes por el valor constante proveniente de una fuente externa es más válida que la sustitución por la media (valor generado internamente).

En el *método de imputación por regresión* se utiliza el análisis de la regresión para predecir los valores ausentes de una variable basándose en su relación con otras variables del conjunto de datos a partir de la ecuación de regresión que las liga. Como desventaja de este método destacaríamos que refuerza las relaciones ya existentes en los datos de modo que conforme aumente su uso los datos resultantes son más característicos de la muestra y menos generalizables. Además, con este método se subestima la varianza de la distribución. Y no olvidemos como desventaja que este método supone que la variable con datos ausentes tiene correlaciones sustanciales con otras variables.

El *método de imputación múltiple* es una combinación de varios métodos de entre los ya citados.

4.2. TÉCNICAS DE DETECCIÓN DE VALORES EXTREMOS

Un valor *outlier* o atípico es una puntuación extrema dentro de una variable. Este tipo de valores afecta fuertemente a los análisis en que inter-

venga la citada variable, sobre todo si trabajamos con muestras pequeñas. Por ejemplo, si estamos trabajando con un modelo de regresión lineal en el que interviene la variable, la distorsión producida normalmente es el sesgo, hacia arriba o hacia abajo, del valor del coeficiente estimado.

Más concretamente, podemos definir los valores atípicos como observaciones aisladas cuyo comportamiento se diferencia claramente del comportamiento medio del resto de las observaciones. Existe una primera categoría de casos atípicos formada por aquellas observaciones que provienen de un error de procedimiento, como por ejemplo un error de codificación, error de entrada de datos, etc. A veces estos datos pueden ser detectados analizando si están comprendidos entre determinados valores plausibles. Por ejemplo si estudiamos la edad de los alumnos de primaria que reciben becas es lógico pensar que su edad estará comprendida digamos entre 5 y 13-14 años (contando con la presencia de repetidores) si en la base de datos tenemos observaciones con edades de 66, 67 o 86 años es lógico pensar que se puede haber producido un error en el proceso de trasvase de datos. Estos errores son frecuentes en el trabajo aplicado y de ahí la importancia de realizar las estadísticas descriptivas vistas en el capítulo 3 de forma previa a cualquier análisis posterior. Estos datos atípicos, si no se detectan mediante filtrado, deben eliminarse o recodificarse como datos ausentes.

Otra categoría de casos atípicos contempla aquellas observaciones que ocurren como consecuencia de un acontecimiento extraordinario existiendo una explicación para su presencia en la muestra. Este tipo de casos atípicos normalmente se retienen en la muestra, salvo que su significancia sea sólo anecdótica.

Otra categoría adicional de datos atípicos comprende las observaciones extraordinarias para las que el investigador no tiene explicación. Normalmente estos datos atípicos se eliminan del análisis. Una última categoría de datos atípicos la forman las observaciones que se sitúan fuera del rango ordinario de valores de la variable. Suelen denominarse valores extremos y se eliminan del análisis si se observa que no son elementos significativos para la población. Por ejemplo, al estudiar la influencia del número de hijos sobre determinadas variables económicas podremos ver que la mayoría de familias tendrán un número de hijos comprendido entre digamos 0 y 5. Pero no es imposible que existieran familias con 10 hijos, cuyo análisis debería ser tratado de forma cautelosa ya que no es representativo de la realidad social a estudiar si no queremos que este dato afecte a los análisis que hagamos posteriormente.

Las propias características del caso atípico, así como los objetivos del análisis que se realiza, determinan los casos atípicos a eliminar. No obstante, los casos atípicos deben considerarse en el conjunto de todas las varia-

bles consideradas. Por lo tanto, hay que analizarlos desde una perspectiva multivariante. Puede ocurrir que una variable tenga valores extremos eliminables, pero al considerar un número suficiente de otras variables en el análisis, el investigador puede decidir no eliminarlos.

Pueden utilizarse *herramientas de análisis exploratorio de datos* para *detectar casos atípicos en un contexto univariante*. Por ejemplo, en el gráfico de caja y bigotes los valores atípicos se presentan como puntos aislados en los extremos de los bigotes. Los valores extremos suelen aparecer tachados con una *x*. El *software* habitual indica el número de observación correspondiente a los valores atípicos. En la Figura 4-5 se muestra el gráfico de caja y bigotes para una variable. Se observan varios valores atípicos por encima del bigote superior y otros por debajo del bigote inferior. La utilidad de XLSTAT permite también representar los valores máximo y mínimo.

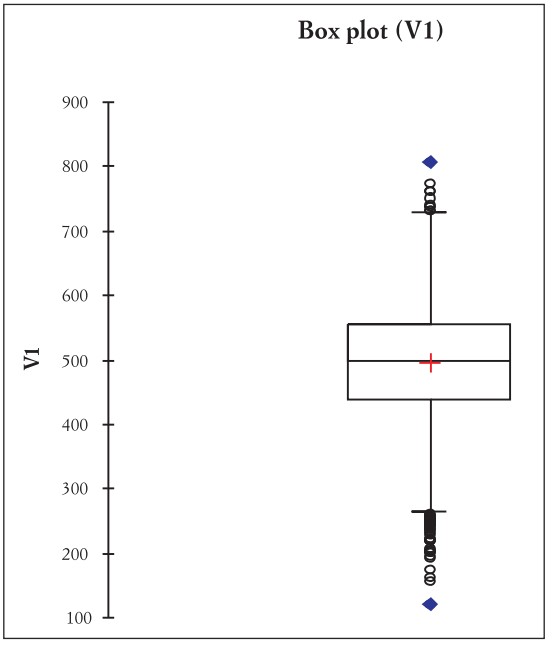


FIGURA 4-5

Otro camino para detectar valores atípicos consiste en utilizar un *diagrama de control*, consistente en una representación gráfica con una línea central que denota el valor medio de la variable y con otras dos líneas horizontales, llamadas *Límite Superior de Control* (LSC) y *Límite Inferior de Control* (LIC). Se escogen estos límites de manera que casi la totalidad de los puntos de la variable se halle entre ellos. Mientras los valores de la variable se encuentran entre los límites de control, se considera que no hay

valores atípicos. Sin embargo, un punto que se encuentra fuera de los límites de control se interpreta como un valor atípico, y son necesarias acciones de investigación y corrección a fin de encontrar y eliminar la o las causas asignables a este comportamiento.

Se acostumbra a unir los diferentes puntos en el diagrama de control mediante segmentos rectilíneos con objeto de visualizar mejor la evolución de la secuencia de los valores de la variable.

Sin importar la distribución de la variable, es práctica estándar situar los límites de control como un múltiplo de la desviación típica. Se escoge en general el múltiplo 3, es decir, se acostumbra a utilizar los *límites de control de tres sigmas* en los diagramas de control. En excel podemos establecer los límites de control mediante funciones lógicas.

Se utilizan límites de tres sigmas porque la mayoría de las distribuciones con que nos encontramos en la práctica se aproximan a la forma de campana de Gauss (función de densidad de la distribución normal).

Como indica la Figura 4-6, la probabilidad de encontrar un valor dentro de $\mu \pm \sigma$ es aproximadamente del 68%. Similarmente, la probabilidad de que los valores caigan fuera de los límites $\mu \pm 2\sigma$ es aproximadamente del 4,5%, mientras que la probabilidad de que los valores caigan fuera de los límites $\mu \pm 3\sigma$ es despreciable (sólo del 0,3% o del tres por mil). Por esta razón se utilizan límites tres sigmas.

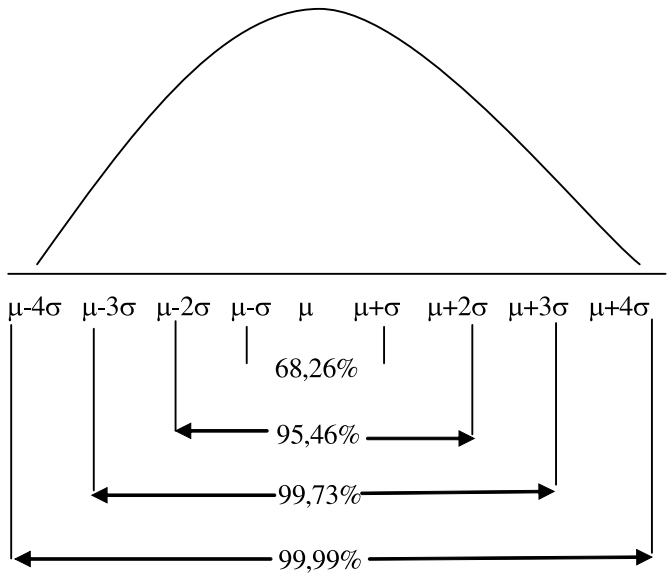


FIGURA 4-6

También se detectan posibles valores atípicos mediante los *estadísticos robustos de la variable* y ver su diferencia respecto de los estadísticos no robustos. Suelen considerarse como estadísticos robustos de centralización (localización) la mediana, la media truncada y la media winsorizada. La media truncada prescinde del 15% de los valores de la variable por cada extremo y la media winsorizada sustituye ese 15% de valores por valores del centro de la distribución. Como estadísticos robustos de dispersión (escala) se usan respectivamente la variación media respecto de la mediana, la desviación típica truncada y la desviación típica winsorizada. Cuando no hay valores atípicos, los estadísticos robustos y los estadísticos normales no difieren mucho. También pueden calcularse intervalos de confianza para la media normal y para la media winsorizada. Si su anchura es similar no hay valores atípicos.

Cuando se trata de *detectar casos atípicos en un contexto bivariante*, pueden utilizarse *herramientas de análisis exploratorio de datos*, por ejemplo, el gráfico de caja y bigotes múltiple (Figura 4-7) que representa distintos gráficos de una variable (resultado en matemáticas) para diferentes niveles de la otra (tipo de familia en la que vive: nuclear y otra).

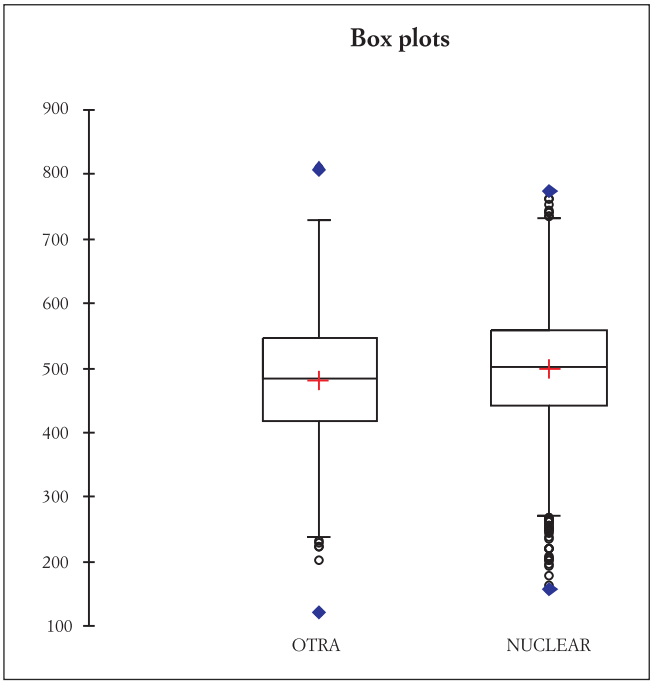


FIGURA 4-7

Otra forma de detectar casos atípicos en un contexto bivariante consiste en evaluar conjuntamente pares de variables mediante un gráfico de dispersión. En la Figura 4-8, que representa el resultado en matemáticas en función de su nivel económico, aparecen casos que caen manifiestamente fuera del rango del resto de las observaciones y por tanto pueden identificarse como puntos aislados en el gráfico de dispersión.

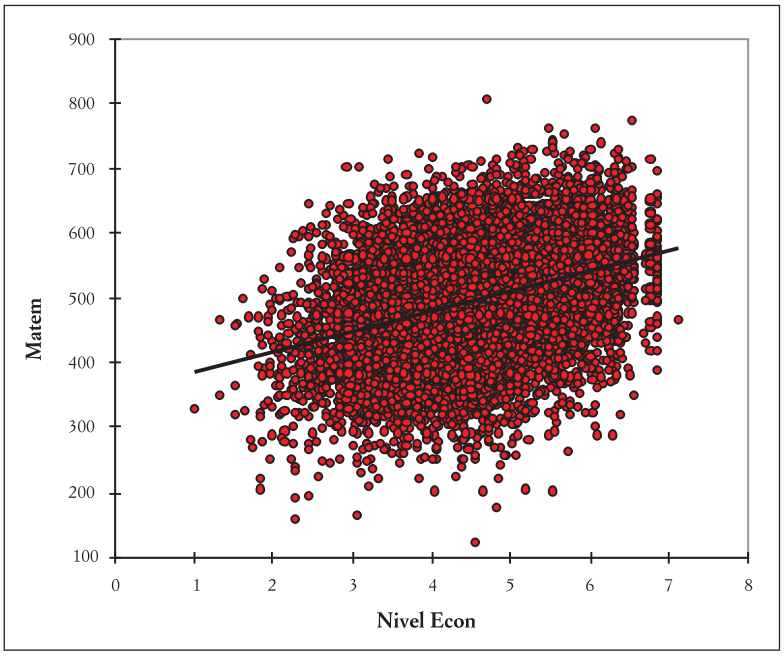


FIGURA 4-8

EJERCICIO 4-1. El fichero PISA2003ESPAÑA.XLS contiene los resultados de los alumnos españoles de 15 años que participaron en la prueba PISA 2003. Dado que hay información faltante en alguna de las variables, se trata de: 1) detectar que variables presentan datos perdidos, 2) reemplazar los datos perdidos utilizando el método de imputación de la mediana. 3) En caso positivo realizar la imputación de los datos ausentes.

Para realizar un análisis de valores perdidos utilice la función CONTAR.BLANCO(Celda inicial; celda final) para cada una de las variables. De esta forma podrá comprobar que las variables NUCLEAR y MONOPARENTAL presentan 63 casos perdidos mientras que la variable PREESCOLAR presenta 123 casos perdidos. Calcular

Tras observar la presencia de datos ausentes en una distribución será necesario detectar si éstos se distribuyen aleatoriamente. Habrá que detectar que el efecto de los datos ausentes es importante mediante pruebas formales de aleatoriedad.

Ya sabemos que una *primera prueba para valorar los datos ausentes* es la *prueba de las correlaciones dicotomizadas*. Para realizar esta prueba, para cada variable V del análisis se construye una variable dicotomizada asignando el valor cero a los valores ausentes y el valor uno a los valores presentes. A continuación se halla la matriz de correlaciones de las variables cuantitativas dicotomizadas acompañada de los contrastes de significatividad de cada coeficiente de correlación de la matriz. Si los elementos de la matriz de correlaciones son no significativos, los datos ausentes son completamente aleatorios. Si existe alguna correlación significativa y la mayor parte son no significativas, los datos ausentes pueden considerarse aleatorios. En ambos casos podrán realizarse análisis estadísticos previa imputación de la información faltante.

Comenzamos generando las variables D1 a D2 (D1 corresponde a la variable NUCLEAR y D2 a PREESCOLAR de modo que *Di* vale 0 para valores ausentes y *Di,* vale 1 para valores presentes. Dado que los datos perdidos figuran como «blancos» es conveniente asignar previamente a los cuadros en blanco un valor como 999 de tal forma que las casillas con este valor sabremos que son datos perdidos. Para ello utilizamos *Edición* → *Reemplazar.* Posteriormente utilizando la función lógica SI daremos valor 1 a los valores distintos de 999 y 0 a los datos válidos (o viceversa)

Para obtener la matriz de correlaciones bivariadas elegimos en los menús *Herramientas* → *Análisis de Datos* → *Coeficiente de Correlación* y seleccionamos las variables numéricas D1 y D2. Al pulsar en *Aceptar*, se obtiene la matriz de correlaciones. En esta matriz el coeficiente de correlación tiene un valor de 0,014 por lo que puede considerarse que los datos ausentes se distribuyen aleatoriamente. Podemos entonces aplicar técnicas de imputación de la información faltante.

Una *segunda prueba para valorar los datos ausentes* para una única variable Y consiste en formar dos grupos de valores para Y, los que tienen datos ausentes y los que no los tienen. A continuación, para cada variable X distinta de Y, se realiza un test para determinar si existen diferencias significativas entre lo dos grupos de valores determinados por la variable Y (ausentes y no ausentes) sobre X.

Si vamos considerando como Y cada una de las variables del análisis y repitiendo el proceso anterior se encuentra que todas las diferencias son no significativas, se puede concluir que los datos ausentes obedecen a un *proceso completamente aleatorio* y por tanto pueden realizarse análisis estadísticos

fiables con nuestras variables *imputando los datos ausentes* por los métodos que se verán más adelante. Si un porcentaje bastante alto de las diferencias son no significativas, puede considerarse que los datos ausentes obedecen a un *proceso aleatorio* (no completamente aleatorio) que también permitirá realizar análisis estadísticos fiables con nuestras variables previa *imputación de la información faltante*, aunque con menos fiabilidad que en el caso anterior.

Una vez que se ha contrastado la existencia de aleatoriedad en los datos ausentes ya se puede tomar una decisión para dichos datos antes de comenzar cualquier análisis estadístico con ellos. En este ejercicio podemos comenzar incluyendo sólo en el análisis las observaciones (casos) con datos completos (filas cuyos valores para todas las variables sean válidos), es decir, cualquier fila que tenga algún dato desaparecido se elimina del conjunto de datos antes de realizar el análisis. Este método se denomina *aproximación de casos completos* o *supresión de casos según lista* y suele ser el método por defecto en la mayoría del software estadístico. Este método es apropiado cuando no hay demasiados valores perdidos, como en nuestro caso.

EJERCICIO 4-2. Para los datos del problema anterior realizar un análisis de la presencia de datos atípicos en las variables matemáticas, lectura y ciencias.

Para el *análisis de los datos atípicos* se puede utilizar el diagrama de caja y bigotes. Para ello, puede utilizarse *XLSTAT → Visualización de Datos → Gráficos Univariados* (Figura 4-9).

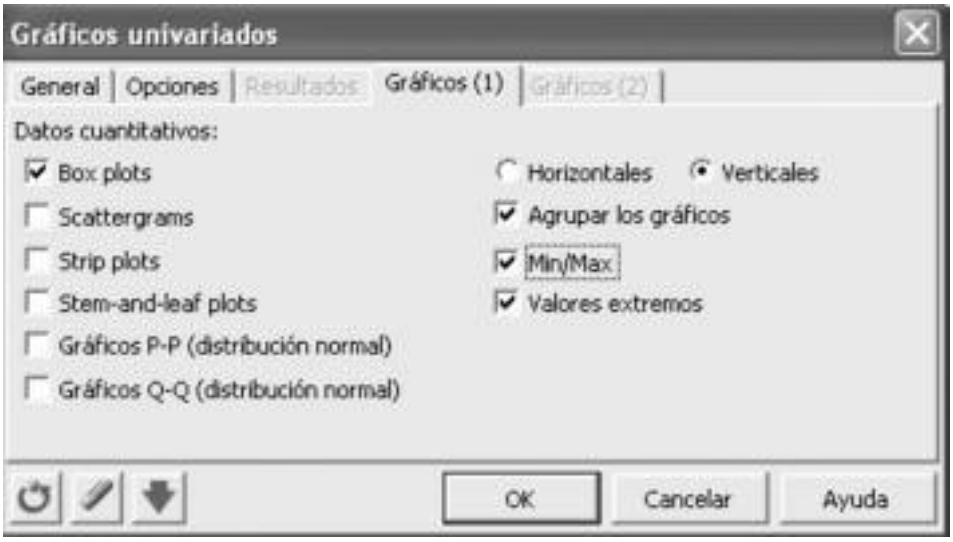


FIGURA 4-9

Seleccionamos la opción box-plot, agrupar los gráficos (para obtener en un solo gráfico las tres variables), Min/Max y valores extremos. Al pulsar en *Aceptar* se obtiene el gráfico de caja y bigotes (figura 4-10) para cada una de las variables. Según este gráfico hay valores atípicos en las tres variables fundamentalmente en ciencias.

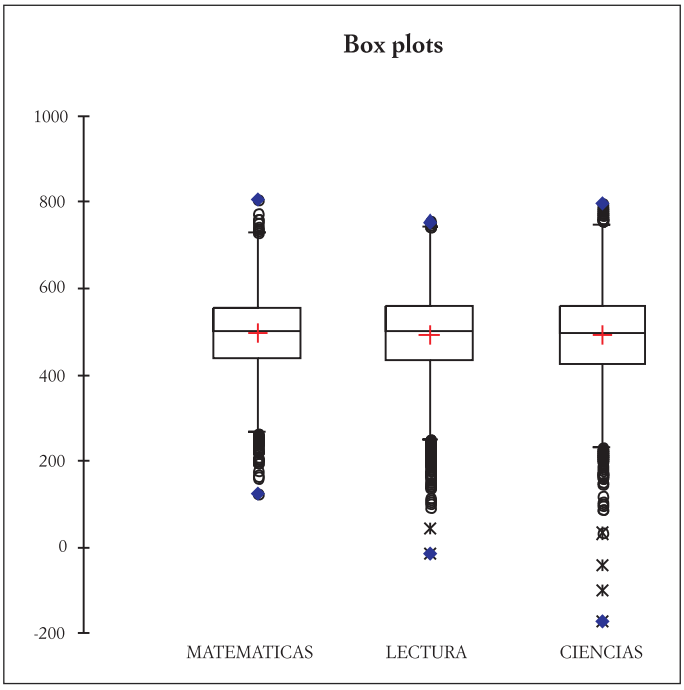


FIGURA 4-10

BIBLIOGRAFÍA

Pérez López, C. (2005). Métodos estadísticos avanzados con SPSS. Thomson Paraninfo.

Pérez López, C. (2005). Técnicas estadísticas con SPSS12. Aplicaciones al análisis de datos. Pearson Alambra.

Pérez López, C. (2004). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS. Pearson Alhambra.

CAPÍTULO V
INDICADORES, EFICACIA, EFICIENCIA Y NECESIDAD DE EVALUACIÓN

AURELIA VALIÑO CASTRO

5.1. NECESIDAD DE EVALUACIÓN DE LAS POLÍTICAS DE FAMILIA

En los últimos veinte años se han producido importantes cambios en las características económicas y sociales de las familias en España, que plantean nuevas necesidades y se añaden a los problemas de falta de atención que tradicionalmente han tenido en otros campos.

Es imprescindible efectuar una evaluación de las necesidades que surgen a tenor de estos cambios y de las políticas que se emplean para ayudar a satisfacerlas, comprobando el alcance de objetivos y eficacia de las distintas alternativas empleadas. Toda política exige una evaluación de este tipo, de hecho el campo de las políticas públicas en general tiene unas peculiaridades que exigen un tratamiento específico, pero en el caso de las políticas de familia reviste una especial importancia por la falta de tradición en su análisis.

Pretendemos aquí examinar las características de las políticas de familia que condicionan la evaluación de las mismas y a la inversa, las posibilidades de aplicación de las técnicas de evaluación disponibles a las políticas de familia. El avance en el campo de la informática y la extensión de su utilización a todos los ámbitos ha permitido y favorecido la posibilidad de acceso a grandes bases de datos y tratamiento de información que antes quedaba circunscrito a un grupo muy específico de especialistas. Pero en ocasiones también se produce un abuso con tratamientos innecesariamente sofisticados que no aportan resultados relevantes de cara a la valoración. Pretendemos clarificar las posibilidades en este campo delimitando los métodos en cuanto a su utilidad y viabilidad de aplicación al campo concreto de las políticas de familia.

Efectuamos un repaso a modo de presentación de las posibilidades de evaluación que brindan las políticas públicas, ofreciendo una panorámica general. Algunas de ellas se desarrollan en este libro con detalle, pero sería imposible reunir todas. La imaginación del evaluador tiene un campo totalmente disponible para inventar las mecánicas mas adaptadas al problema que tiene que resolver. Sólo le presentamos

algunas ideas que le pueden producir otras incluso más interesantes. Tampoco pretendemos explicar todas las técnicas al detalle, no sólo excedería las dimensiones impuestas a este capítulo, sino al propio objetivo de la obra. El lector que pretenda profundizar en alguna de las técnicas que se presentan deberá acudir a bibliografía específica, alguna de la mas relevante o sencilla se recoge al final del capítulo en la bibliografía adjunta. De entre todos los sistemas alternativos o concurrentes de evaluación se va a prestar especial atención en este capítulo a la evaluación por objetivos a través del uso de indicadores. Las razones son múltiples, pero destacan especialmente: que este es un método que permite resaltar las peculiaridades de las políticas a las que se aplica; que la información que se extrae de su aplicación y los indicadores como tales son la base y punto de partida para otros métodos más complejos; que los indicadores se pueden calcular con mayor rapidez y se pueden automatizar los cálculos; que su sencillez permite el uso de los mismos por parte de los no iniciados en técnicas econométricas de relativa complejidad; y que, una vez establecidos, permiten efectuar un seguimiento en el tiempo de las políticas evaluadas. Pero a pesar de su sencillez se puede obtener información relevante. De hecho se puede obtener un diagnóstico de la situación y necesidades de las familias, pueden ayudar a establecer sistemas de valoración que pongan de relieve los beneficios sociales alcanzados por las políticas aplicadas y, por lo tanto, la ratificación de las mismas o, por el contrario, su insuficiencia para alcanzar los objetivos propuestos y, por lo tanto, la necesidad de reformarlas.

Características de la evaluación de las políticas de familia

Las políticas de familia se enfrentan a los problemas de evaluación generales de toda medida pública resumidos en la ausencia de las condiciones de mercado que impiden el reflejo en los precios de las valoraciones de los ciudadanos. Este problema, junto con los derivados de la necesidad de redistribuir renta, sería el núcleo básico del que derivan otros también específicos del Sector Público, tales como: diversidad de actividades, tamaño excesivo, ambigüedad en la tecnología (dificultades para conocer las funciones de producción, o como interfieren y en qué medida distintas variables para obtener un resultado en otra variable), diferentes niveles de administración compartiendo competencias, menor experiencia en la evaluación que el sector privado, problemas en la especificación de objetivos y en la determinación de la población objetivo. Todas estas características dan lugar a la necesidad de técnicas de evaluación específicas o adaptadas al Sector Público. La aplicación de las mismas a cada una de las políticas públicas exige también un tratamiento diferenciado.

Concurrencia de administraciones

Uno de los aspectos más destacados, entre los antes enumerados, en las políticas de familia es la concurrencia de administraciones. Los países tienen distintos niveles de gobierno con competencia para dictar normas y dotar ayudas de distinto tipo a una misma familia que pueden tanto añadir como restar de cara a un resultado final o a un objetivo de una administración concreta. Por ejemplo, en España las políticas de familia se ven afectadas por normativa de la Unión Europea, que dicta directrices, resoluciones, consejos y recomendaciones referentes a políticas de familia o que afectan a la familia; por leyes del gobierno central que regulan el tratamiento de la familia [1] y sus miembros en el ámbito nacional; por prestaciones y ayudas de las Comunidades Autónomas y de los Ayuntamientos. Uno de los principales problemas de esta concurrencia de órganos de decisión, actuando sobre la misma población objetivo (la familia) y en algunos casos intentado conseguir los mismos objetivos, de cara a evaluar una política pública es que dificulta en gran medida la diferenciación de los efectos finales que cada actuación de cada nivel administrativo pueda tener.

No tenemos una regla de oro que indique cuales son los límites que en teoría son más aconsejables para cada administración. Son las normas que regulan las competencias las que al final delimitan la solución final y que suelen tener en cuenta cuándo y cuánto favorece la cercanía al usuario final a la hora de decidir y gestionar los servicios a prestar. El grado de descentralización de la asistencia social, tan relacionada con las ayudas a la familia, no es uniforme a nivel internacional. Entre los sistemas altamente centralizados se encuentran Australia, Dinamarca, la República Checa, Japón, Luxemburgo, Méjico, Nueva Zelanda, Portugal, Grecia, Irlanda, Turquía, Reino Unido. Entre los altamente descentralizados se encuentran Austria, Canadá, Italia, Noruega, Suiza. Y entre los que están en una posición intermedia: Bélgica, Finlandia, Alemania, Suecia, Estados Unidos, Polonia, Hungría. Hasta épocas recientes España se encontraba en una posición intermedia, pero en los últimos años se ha unido a los altamente descentralizados. Los centralizados establecen las condiciones y características y financian, en general, las prestaciones asistenciales, encargándose las autoridades descentralizadas en muchos casos de la gestión, por lo que una característica inmediata es la uniformidad en las prestaciones. En los altamente descentralizados la regulación nacional suele limitarse a establecer los ámbitos a cubrir, por lo que las diferencias son muy importantes. Y los sistemas intermedios tienen normas nacionales que establecen características comunes, pero los gobiernos locales o regionales tienen capacidad para establecer ayudas complementarias que establecen diferencias territoriales y que se financian a través de transferencias de los gobier-

[1] Tenemos así regulación europea que afecta a cuestiones tales como la conciliación de la vida familiar y laboral, tratamiento de minusválidos, o actividades para la inclusión social.

nos centrales e impuestos propios [2]. A modo de ejemplo, en España las Comunidades Autónomas prestan ayudas a las familias por nacimiento de hijos o adopción, para pedir excedencias por cuidado de hijos, por hijos menores de 3 años, por familia numerosa, por partos múltiples, por hijos que viven fuera del hogar realizando estudios universitarios, por guarderías, por empleados cuidando de hijos menores de 3 años, o amas de casa que no trabajan fuera, bonificaciones en matrículas de universidad y / o compra de libros, carné descuento en lista de establecimientos, ayudas en vales de leche, transporte o ayudas de diverso tipo a través de la vivienda y a las que habría que añadir ayudas a rentas mínimas o «salarios de integración». Estas ayudas no se prestan ni mucho menos en todas las Comunidades Autónomas y cuando coinciden en la aplicación de alguna, las características (monetarias o en especie, gastos directos o fiscales), los requisitos y cuantías son muy diversos. Así pues existen profundas diferencias entre las familias en función de su residencia. Las ayudas prestadas por los ayuntamientos van más dirigidas a ayudas en guarderías y ayudas asistenciales a mayores.

Las políticas de familia a su vez se relacionan, entrecruzan o superponen con los objetivos de diversos órganos dentro del mismo nivel competencial (ministerios, consejerías o concejalías). Como hemos visto, las políticas de familia afectan a campos tan diversos como la política laboral, política educativa, atención a menores, atención a los ancianos, regulaciones sobre malos tratos, adopciones, política de vivienda o políticas de prevención o ayuda frente a la exclusión social, por citar algunos ejemplos de políticas conexionadas con las que pudiéramos llamar estrictamente de familia, como las de ayuda a familias numerosas o ayudas de conciliación de vida familiar y laboral. Estas conexiones obligan a que las Administraciones colaboren entre sí por el bien de las familias y sus propias políticas, a fin de evitar interferencias que pudieran perjudicar la obtención de algunos objetivos.

Concurrencia o interferencia de objetivos

La existencia de interrelaciones entre los objetivos se ha de tener presente a la hora del diseño y posterior evaluación de las políticas públicas. Como ejemplo destacan las concurrencias entre las ayudas a la familia, las ayudas a la igualdad de hombres y mujeres y ayudas a la infancia, así como ayudas a la tercera edad. Y uno de los aspectos donde pueden concurrir estas ayudas, ya sea de forma conflictiva o favorecedora, es en la búsqueda de «conciliación de la vida familiar y laboral».

---

[2] Tomado de M.ª Teresa López y Aurelia Valiño (2000) Análisis de las Políticas de la Exclusión social en la Comunidad de Madrid. Investigación realizada para la Comunidad de Madrid exp. Nº 06/0120/1999, citando a Kalisch, D.W.; Aman, T. Y Buchele, L.A. (1998) Social and Health Policies in OECD countries: a survey of current programmes and recent developments. Labour market and social policy. Occasional papers nº 33.

Son clásicas también las discusiones doctrinales en torno a los conflictos entre los objetivos de eficiencia y de equidad. Estos conflictos pueden superarse si los objetivos de equidad se introducen en los objetivos de eficacia, y de alguna manera en los sistemas de evaluación, como veremos más adelante. De hecho, alguna de las medidas de eficacia y eficiencia incorporan sistemas para medir la equidad. Pero en ocasiones la equidad es un objetivo en sí mismo, y dispone de sistemas o técnicas de evaluación propias y específicas.

Conseguir un cierto grado de equidad implica actuar sobre las distintas dimensiones que presenta este concepto dependiendo del objetivo que se pretenda conseguir. Así podemos concebir la equidad como:

- *Igualdad de oportunidades* que representa una noción de equidad ex-ante donde tienen cabida todas aquellas medidas que intentan proporcionar al individuo unas similares condiciones de partida (en este caso tendríamos intervenciones públicas encaminadas a garantizar igual acceso a la educación, sanidad, justicia, etc.).

- Muy relacionada con la anterior noción esta la *equidad categórica*, considerada ésta como aquella situación en la que todos los individuos tienen derecho y, en ocasiones, están obligados a consumir determinados bienes denominados *Sociales* (educación, sanidad, vivienda, cultura, etc.).

- *Equidad redistributiva*, o equidad ex post, cuyo objetivo es conseguir a través de los programas sociales una disminución de las desigualdades de renta y riqueza de los ciudadanos. Bien es cierto, que no todos los programas tienen la misma capacidad redistributiva, dependerá del objetivo que se persiga, del carácter de la medida introducida y de la población a quien va dirigida.

En este sentido, podemos decir que las políticas públicas de apoyo a la familia son una parte del conjunto de la política global de protección social y pueden ser analizadas desde muy diversas perspectivas, dependiendo del objetivo concreto que cada una de ellas pretenda conseguir en términos de equidad. La efectividad de una medida requiere del establecimiento de una serie de indicadores que nos permitan hacer una evaluación y un seguimiento continuado de su incidencia en términos de equidad.

5.2. MÉTODOS DE EVALUACIÓN

En general la evaluación pretende determinar la relevancia, eficiencia y eficacia de una actividad en la consecución de unos objetivos, tan sistemáti-

camente y objetivamente como sea posible, incluyendo un análisis de la aplicación y gestión administrativa de tal actividad. (Papacostantinou, G y Polt, W. 1997, p. 10) [3].

A la hora de decidir el instrumento o instrumentos de valoración más adecuados hay que determinar primero cual es el objetivo de la evaluación, cuando va a tener lugar, la información disponible y quien evalúa. Es decir, habrá que dar respuesta a las preguntas de ¿por qué?, ¿quién?, ¿cuándo? y ¿cómo?. En la evaluación están interesados los políticos, los gestores, los analistas económicos y los ciudadanos. Las razones de la evaluación son múltiples y específicas para cada uno de los anteriores, pero en general podrían resumirse en:

- sopesar la justificación de un determinado programa de actuación,
- analizar sus efectos económicos especialmente en la población a la que va dirigido,
- obtener información para la adecuada gestión de los recursos y en relación con esto conseguir mejorar la calidad en los servicios prestados,
- y escoger el mejor método o medida de actuación en función de su menor coste, y máxima eficacia, eficiencia y equidad.

Sobre las técnicas de evaluación se pueden hacer distintas clasificaciones. Nosotros vamos a recoger las más relevantes que se basan en presentar los métodos más útiles para el diseño de las políticas (o valoración «ex ante»), para el seguimiento de las mismas (o valoración «durante»), o para la comprobación de sus resultados (o valoración «ex post»). A través del desarrollo de estos métodos se estaría dando respuesta a las tres primeras preguntas de ¿por qué?, ¿quién? y ¿cuándo? ; y otras clasificaciones, como las que diferencian en técnicas cualitativas o cuantitativas, estarían dando respuesta a la última pregunta de ¿cómo?. A su vez cada medida o instrumento de política concreta aplicado exige su propia metodología. No es lo mismo una medida realizada a través de regulación, de un gasto fiscal o de un gasto directo [4].

---

[3] Papaconstantinou G. and Polt W. (eds) (1997). Policy Evaluation in Innovation and Technology- Towards Best Practices, OECD.

[4] Son múltiples las referencias que podríamos presentar para el análisis de políticas públicas y métodos de valoración, que además se multiplican por cada uno de los métodos que vamos a presentar. No disponemos, sin embargo de una referencia general para métodos de evaluación en políticas de familia, aunque sí tengamos para la evaluación de alguna de las políticas de familia. Como referencia para el enfoque general de la evaluación se puede citar: Eugene Bardach (2000): *A Paractical Guide for Policy Análisis,* New York. NY: Chatham House Publishers.

Otra clasificación es la que atiende a los enfoques básicos de la evaluación:

- *Economía o análisis financiero*, que pone en relación los costes con los recursos. Responde a cuanto cuestan los recursos disponibles para aplicar la política en cuestión.

- *Eficiencia o análisis de productividad* que busca determinar las medidas que permiten obtener la máxima producción a un coste dado o minimizar el coste de una producción dada. Pone pues en relación los recursos utilizados con la producción obtenida (output). Se encuentra muy relacionada con los niveles de servicio prestados y el grado en el que están siendo utilizados por la población objetivo (aquella a la que van dirigidas las políticas)

- *Eficacia o análisis de resultados* que busca medir en cuanto se alcanzan los objetivos propuestos. Los resultados (outcomes) se establecen en función de los efectos obtenidos sobre la población objetivo (impacto). Para que las medidas adoptadas sean eficaces los resultados obtenidos deben acercarse suficientemente a los propuestos. Como antes hemos señalado los objetivos de equidad se incorporan al análisis considerándolos como un objetivo más de las políticas analizadas.

Estos enfoques se presentan tanto en la valoración para el diseño de políticas, en el seguimiento de las mismas o en la evaluación final.

Una conclusión que podemos adelantar ya es que no hay una solución única o una regla metodológica única. Cada actuación puede tener su propia metodología de evaluación. Nosotros vamos a presentar, con la mayor amplitud posible, toda la batería de posibilidades de evaluación, especificando, cuando sea conveniente, el destino de política concreta de familia a la que se debe aplicar. En algunos casos, incluso lo más adecuado es utilizar varias técnicas.

Valoración ex-ante o de diseño de políticas públicas

La valoración ex-ante o prospectiva va dirigida al diseño de políticas, a establecer cuales son los problemas a resolver u objetivos a atender y a escoger el sistema más adecuado para resolverlos. Esta última elección va a estar en función del coste de dicho sistema, y de si sirve para lo que se pretendía inicialmente.

A la hora de diseñar una política es necesario tener información sobre las necesidades a las que pretende atender y para ello es imprescindi-

ble conocer las características de la población objetivo; es decir, de la población a la que van dirigidas o sobre la que se quiere influir. Esta población está condicionada a su vez por el objetivo que pretenda alcanzarse con la política en cuestión. Así por ejemplo, si el objetivo es aumentar el bienestar de las familias, la población objetivo será la familia en general, pero si lo que pretende la medida es atender a la natalidad, obviamente el objetivo serán las familias en edad fértil. En general, habrá que conocer las características sociales y económicas de la familia y sus miembros, estableciendo el *estado de situación* de partida a fin de conocer los problemas más relevantes a los que se enfrenta y sus necesidades. Esta información es útil para, una vez aplicado el plan de actuación, comprobar los efectos producidos a través de los cambios experimentados en la situación de partida.

Conocer el *estado de situación* es el punto de partida para cualquier diseño de políticas públicas y por lo tanto para realizar una *valoración ex-ante*. Para establecer el *estado de la situación*, disponemos de técnicas estadísticas simples, basadas en indicadores, con análisis de medias, varianzas, etc. Y de otras también estadísticas, aunque relativamente más complejas, como por ejemplo técnicas de clasificación del tipo *cluster*, también denominadas de *análisis de conglomerados*. Estas técnicas serían muy útiles para agrupar poblaciones en función de características comunes. Así, por ejemplo, podríamos comprobar si hay zonas en la Comunidad de Madrid, lo que ocurre con toda seguridad, en las que la proporción de familias con ascendientes dependientes es mucho mayor, o por el contrario predominan las familias con hijos menores de 6 ó 3 años, etc. Esto es relevante de cara a poder discriminar la atención, por tipos de necesidades, que ha de presentarse en cooperación con los Ayuntamientos; o incluso dentro de los Ayuntamientos diferenciar también el tipo de necesidad predominante por áreas de distrito. Información ésta que es imprescindible para optimizar la asignación de recursos escasos. Para poder realizarlo es necesario contar con la información de las características sociales y económicas de las familias por zonas geográficas, en función de la agrupación que se pretenda realizar. La aplicación intrínseca del *análisis de conglomerados* es clasificar y etiquetar, y es imprescindible conocer, clasificar y etiquetar a la población objetivo y sus necesidades para aplicar una política coherente con las mismas [5].

En esta misma línea de segmentación podrían utilizarse técnicas de análisis discriminante o mejor aún, por su mayor visualización inmediata y por lo tanto comprensión, árboles de decisión. El análisis discriminante permite estudiar las diferencias entre dos o más grupos de individuos previamente definidos con respecto a varias variables simultáneamente y ex-

[5] Un ejemplo de la aplicación del método de cluster a políticas de familia puede verse en el documento: Valiño A. y López .T (2006) Factores explicativos de las ayudas directas a las familias por Comunidades Autónomas. Documento nº 2006-005. Facultad de Ciencias Económicas y Empresariales http://www.ucm.es/BUCM/cee/050501.htm#13.

plicar en qué sentido se dan y así, cuando dispongamos de una nueva observación, poder incluirla en alguno de los grupos seleccionados. Por ejemplo, podríamos estudiar las diferentes necesidades que se plantean para las familias en función de su composición: sin hijos, con 1 hijo, con 2 hijos, con 3 hijos, con 4 o más; o entre las familias con hijos menores de 12 y mayores dependientes; o cualquier otra agrupación que consideremos interesante para nuestro análisis y que pudiera dar lugar a una clasificación interesante para *discriminar* las políticas públicas. Como su nombre indica, los *árboles de decisión* son técnicas de clasificación que ayudan a decidir. También son modelos de predicción. Serían útiles para diseñar, por ejemplo, las medidas de ayuda a las familias con riesgo de exclusión social. En primer lugar nos indicaría la proporción de familias que se encuentra en cada grupo y la probabilidad de que caigan en la situación de exclusión social una vez que ésta está definida. Podría ayudar a definir los requisitos para poder acceder a las ayudas, siendo más o menos restrictivos en función del volumen de presupuesto disponible y el porcentaje de familias que caen dentro del grupo. Además, una vez diseñados son útiles para ayudar a elegir entre diferentes alternativas de políticas a aplicar. Para poder realizar este análisis es necesario añadir al conocimiento de las características económicas y sociales de las familias, aquellas en concreto que afectan a la exclusión social y definir el nivel de renta que la determina, que actuará como variable final a explicar en este ejemplo concreto.

Existen otros métodos que permiten poner unas variables en función de otras y analizar sus relaciones y estaríamos ante métodos estadísticos de dependencia que utilizan técnicas multivariantes, analíticas o inferenciales. Dependiendo del tipo de variables se aplican modelos de regresión múltiple, análisis canónico, análisis discriminante, modelos de elección discreta —logit, probit- modelos ANOVA, ANCOVA, MANOVA, MANCOVA, segmentación jerárquica, etc. Una de las condiciones imprescindibles para que estas técnicas den resultados que sean significativos y puedan utilizarse para la toma de decisiones es disponer de un número de datos suficientemente elevado, condicionado a su vez por el número de variables utilizadas en el modelo. En ausencia de datos suficientes es preferible utilizar sistemas menos sofisticados y más significativos, a fin de evitar llegar a conclusiones erróneas.

Una vez efectuado el diagnóstico de los problemas y necesidades de las familias, el diseño de políticas exige determinar las más adecuadas. Como técnica de selección de proyectos destaca el *análisis Coste- Beneficio*. Esta técnica es especialmente útil para escoger entre varias alternativas aquella que proporciona una mayor rentabilidad social neta. Se compara el valor actual de la corriente futura de beneficios privados y sociales con el valor actual de la corriente futura de costes. Los beneficios y costes sociales han de calcularse a través de métodos indirectos que permitan de-

terminar su valor monetario y ahí radica una de las mayores dificultades del análisis coste- beneficio. A fin de elegir se efectúa una ordenación de medidas en función del mayor valor de beneficios frente a coste, o del mayor ratio de beneficios sobre costes, o la mayor tasa de retorno. Pero el análisis Coste-Beneficio no necesariamente implica la comparación de técnicas alternativas, en ocasiones se utiliza para evaluar una medida concreta como técnica ex-post. Se puede utilizar para evaluar servicios concretos, no para políticas globales y cuando el proyecto es de suficiente envergadura como para justificar el coste que la técnica implica. Esta técnica, y las de su grupo no pueden utilizarse en la evaluación continua, sino que se efectúa para un momento del tiempo. Los análisis de equidad se incorporan al estudio ponderando en mayor medida los efectos que inciden en las familias de menor renta, pudiendo establecer el límite de renta a partir del cual damos más importancia al efecto distributivo.

Cuando las dificultades son tales que resulta imposible convertir los beneficios sociales a valores monetarios se opta por el *análisis Coste- Eficacia*. En este caso se trata de determinar los costes de alcanzar los objetivos propuestos a través de distintas actividades o medidas.

El *análisis Coste-Utilidad* es similar a las anteriores, pero los outputs del programa se miden en términos de utilidad en lugar de medidas monetarias.

Técnicas relacionadas con el *análisis Coste-Beneficio*, porque se suelen utilizar como métodos indirectos para calcular el valor monetario de intangibles, son las técnicas de Valoración Contingente y las técnicas de Valoración Hedónica.

Las técnicas de *Valoración Contingente* se apoyan en encuestas realizadas sobre la población objetivo. Son muy útiles en la etapa de diseño de políticas para determinar el grado de demanda de un servicio potencial y guardan mucha relación con las técnicas experimentales (someter a un experimento de política hipotética a un grupo de referencia). Se aplican también a servicios concretos y no necesitan de sistemas alternativos de aplicación. Serían útiles para valorar servicios de ayuda a tercera edad (teleasistencia, ayuda domiciliaria, etc.) o para servicios similares en general de ayuda al cuidado de dependientes, o cualquier servicio a las familias que implique una prestación directa a las mismas, ya sea con producción pública o con la contratación con el sector privado (provisión pública). Se trataría de preguntar a los potenciales usuarios: ¿cuánto estarían dispuestos a pagar porque se preste el servicio? o ¿cuánto estarían dispuestos a recibir por compensarles de su ausencia?. Frente a la sencillez aparente del método hay toda una mecánica con soporte científico que condiciona la bondad en los resultados. En esta mecánica se demuestra que es muy importante efectuar las dos pregun-

tas y el sistema de encuesta elegido. Como vamos a ver, las *técnicas de Valoración Contingente* se han aplicado también, y en mayor medida, como valoración ex post.

Las técnicas de *Valoración Hedónica* se apoyan en un bien intermedio que tiene un precio de mercado que a su vez depende de un conjunto de características del propio bien y de otros entre los que se encuentra el valor del bien para el que no tenemos mercado y por lo tanto se desconoce su precio. El bien que suele actuar como intermedio es la vivienda, cuyo precio se fija en función de una serie de características de la propia vivienda (metros cuadrados, número de cuartos de baño, calidad de los materiales...) y de características de la zona (barrio de lujo, servicios cercanos, jardines....) y características medioambientales (emisiones de dióxido de carbono, ruidos...). La elasticidad del precio de la vivienda respecto del bien relacionado nos da el valor de este último (como cambia el precio de la vivienda ante cambios en las características del bien relacionado).

$$P_{hi} = Ph\ (S_{i1}, \ldots S_{ij}, N_{i1} \ldots N_{ik}, Q_{i1} \ldots Q_{im})$$

$$\partial P_i / \partial Q_{im} = P(Q_{im})$$

En equilibrio este precio es igual a lo que la gente está dispuesta a pagar (disponibilidad marginal al pago) por una mejora en el intangible. Este sistema es muy útil para valoraciones de bienes medioambientales. La utilidad para políticas de familia es más difícil de encontrar. No hemos querido dejar de presentar esta alternativa ante la posibilidad de utilizarlo recogiendo la posible influencia de los servicios de familia cercanos a la vivienda del usuario y que la imaginación del lector pueda encontrar una aplicación. No obstante consideramos más idóneo el anterior para evaluar los servicios prestados a las familias.

Otra modalidad de análisis ex ante es el *Análisis Multicriterio*. Con este análisis se pretende recoger una característica muy típica de las políticas de familia como es las de tener efectos en la consecución de diversos objetivos, alguno contradictorios entre sí. Se trata de estimar la combinación de estos efectos cuando no tienen una escala común. Intenta evaluar el programa como un todo teniendo en cuenta la combinación de efectos parciales. Esta técnica ayuda a obtener soluciones eficientes. Se alcanza solución cuando no existe una solución alternativa que mejore en uno de los objetivos, sin que empeore, al menos, en otro. Se trata de determinar el conjunto de alternativas eficientes. La ventaja del *Análisis Multicriterio* es que permite actuar sobre las dimensiones múltiples de un problema (sociales, económicos, culturales, éticos, etc.), y comparar los impactos de las alternativas. Los problemas de decisión se representan en una matriz (n x m) donde n es el

número de alternativas y m es el número de criterios de evaluación. Cada columna describe una opción de política y cada fila la realización de las opciones de cada criterio. Puede utilizarse para establecer un orden, para identificar la opción individual más preferida, o para distinguir entre opciones aceptables o no aceptables [6].

Otras evaluaciones utilizan la *técnica del experimento*. Este tipo de evaluaciones pueden servir para comprobar cuales podrían ser los resultados de una medida determinada. Para poder realizarlos es necesario tener un grupo de control sobre el que se aplicaría la política. Por ello esta técnica no es válida cuando las medidas van a ser universales o deberían serlo. Se trata de comprobar, aplicando el programa a un grupo concreto, los efectos que éste tiene comparándolos con otro grupo con idénticas características donde no se aplica. Se trataría de ver si los efectos producidos no ocurrirían si la política no se aplica. A este tipo de análisis se le denomina también *contrafactual*. Al grupo en el que se aplica la política se le denomina grupo *programado o de programa*, y al grupo donde no se aplican, grupo *contrafactual o de control*.

Desde el punto de vista de la equidad en la valoración ex-ante también es necesario realizar un análisis de la situación de partida, a fin de valorar la posición media de las familias en función de sus características sociales y condicionantes económicos y detectar posibles situaciones de riesgo de exclusión, para lo que serían útiles alguno de los indicadores que ya hemos comentado.

Tradicionalmente, el análisis aplicado de la desigualdad ha utilizado un conjunto de instrumentos de medición que nos permiten clasificar distintas distribuciones de renta según la desigualdad que presentan a través de una función que asocia a cada distribución de la renta un número real, que refleja sintéticamente su nivel de desigualdad. Estos índices de desigualdad tienen en cuenta exclusivamente las diferencias o disparidades en los niveles de renta recogidos en un valor que permite cuantificar el resultado y conocer cómo ha evolucionado la desigualdad. Podemos citar entre los indicadores objetivos más utilizados: la desviación media relativa (DMR), la varianza (V), el coeficiente de variación (CV), la desviación típica de los logaritmos (SL), el índice de Gini (G), o la familia de índices de Theil (T).

---

[6] Algunas referencias sobre análisis multicriterio:
Munda, G. (2003) Multicriteria Assessment.
http://www.ecologicaleconomics.org/publica/encyc_entries/Mlticritassess.pdf
Roy B. (1985) - Méthodologie multicritère d'aide à la décision, Economica, Paris.
Roy B. (1996) - Multicriteria methodology for decision analysis, Kluwer, Dordrecht.
Saaty T.L. (1980) - The analytic hierarchy process, McGraw Hill, New York.
Stiglitz J. E. (2002) – New Perspectives on public finance: recent achievements and future challenges, Journal of Public Economics, 86, pp. 341-360.
Vincke Ph. (1992) - Multicriteria decision aid, Wiley, New York.
Zeleny M. (1982) - Multiple criteria decision making, McGraw Hill, New York.

En este análisis del estado de situación de partida, o de evaluación exante, hay enfoques más modernos. En estos, la desigualdad discurre en torno al análisis del bienestar social como concepto más amplio que la desigualdad, lo que supone enmarcar la noción de desigualdad dentro de la denominada Economía del Bienestar. La desigualdad, en este caso, se mide no solamente por los niveles de concentración o dispersión de una distribución de renta si no que tiene en cuenta el valor de la renta media. Los instrumentos habituales que forman parte de la base del análisis para llevar a cabo una comparación del bienestar son Las Curvas de Lorenz Generalizadas y Las Funciones Abreviadas de Bienestar, así como una familia de índices de desigualdad como son: la familia de índices de Atkinson (1970), los coeficientes de Gini generalizados de Donaldson y Weymark (1980,1983) y Yitzhaki (1983) y la familia de índices de entropía generalizada de Theil (1967) y Cowell (1977). Todos los índices incorporan en sus expresiones de forma explícita una preferencia social a través de un parámetro que mide el grado de aversión a la desigualdad que presenta la sociedad.

En este libro se desgranan muchas de estas medidas de equidad aplicadas a la evaluación de las políticas de familia.

Valoración durante o de seguimiento de las políticas públicas

La valoración de seguimiento exige un tratamiento específico, en tanto en cuanto las limitaciones temporales a las que se enfrentan los analistas son mucho mayores que en las valoraciones previas o posteriores. La necesidad de que la información que se deriva de la misma sea oportuna es de primer orden entre las prelaciones que puedan establecerse. El seguimiento continuado exige también una valoración continuada para corregir rápidamente posibles desviaciones y detectar los posibles problemas que pudieran surgir. También es especialmente importante la evaluación de seguimiento no sólo en cuanto a detectar problemas de las propias políticas en sí mismas, sino para detectar fallos en la gestión o cuellos de botella en la prestación de determinados servicios. Por ello adquieren especial relevancia los métodos de valoración que, aunque menos sofisticados, permiten tener una información puntual de la actividad de los gestores y de los resultados inmediatos que se van alcanzando.

La evaluación por indicadores que corresponde a la evaluación por objetivos juega aquí un papel clave. Mas adelante presentaremos con detalle los posibles indicadores para la evaluación de las políticas de familia, deteniéndonos ahora en los aspectos teóricos.

Los indicadores son ratios o cocientes que ponen en relación dos variables, que pueden ser cuantitativas o cualitativas. Estas relaciones deben reflejar: la relevancia del programa, o en qué medida los objetivos del programa

están en relación con las necesidades; lo costoso que puede ser el programa, o relación entre los inputs necesarios con su precio o coste total; la eficiencia o relación entre los bienes o servicios producidos y su coste, o relación entre inputs y outputs; eficacia o relación entre los impactos que tiene el programa y las necesidades de la población a la que se dirigían los objetivos propuestos.

En este análisis, los dos aspectos más relevantes son determinar la población objetivo (aquella a la que va dirigida el proyecto a evaluar) y los objetivos a los que el proyecto sirve. Esto último sin embargo tiene una gran dificultad, ya que en muchos casos los objetivos no están explicitados o no lo están adecuadamente (son vagos o difusos). Son varias las clasificaciones de objetivos. Una posible es la que diferencia entre:

- *Objetivos generales*: los que recogen el impacto de los programas en las necesidades.
- *Objetivos específicos*: representados en términos de resultado o impacto del programa en la población objetivo.
- *Objetivos operacionales*: expresados en términos de outputs o realizaciones.

Otra posible clasificación con la que nos podemos encontrar es la que diferencia entre:

- *Objetivos finales*: variables que recogen el impacto del programa y los resultados.
- *Objetivos intermedios*, que pueden ser:
  - *Variables input*: relacionadas con los recursos materiales y humanos.
  - *Variables operacionales*: quién tiene que realizar qué, cuándo y cómo.
- *Objetivos puente* que asocian inputs y actividades con los resultados finales.

En la práctica no hay un solo indicador que refleje toda la información necesaria para la evaluación, por lo que se suele presentar una batería o conjunto de indicadores. Los indicadores pueden reflejar diversidad de aspectos, así por ejemplo tenemos indicadores de: volumen, productividad, coste, metas temporales, demanda de servicios o disponibilidad de

los servicios. Otra agrupación de indicadores sería aquella que atiende a: necesidades, financieros, actividad- gestión, resultados. La Unión Europea (www.europa.eu.int) distingue entre indicadores de:

- Recursos (financieros)
- Realización (actividad)
- Resultado (efectos directos)
- Impacto (efectos a largo plazo)

Los problemas en la evaluación a través de indicadores surgen cuando las relaciones que se establecen entre las necesidades, los objetivos y la población objetivo, son ambiguas; o cuando las dificultades derivadas de la ausencia de datos obligan al analista a centrarse únicamente en aquellos para los que hay datos disponibles. Por otra parte, los indicadores tienen un carácter estático y fotográfico, aunque este es un problema común con otras técnicas más complejas. Hay que utilizarlos con mucha precaución en las comparaciones. Para conseguir un carácter dinámico sólo sería necesario poder hacer comparaciones a lo largo del tiempo. Esto que parece tan simple se vuelve complejo cuando no se puede mantener una constancia en las bases de datos que se emplean para elaborarlos.

La equidad en este caso, mejor que en ningún otro, se recoge a través de introducir entre las necesidades la de redistribuir renta, y entre los objetivos, el lograr erradicar la exclusión social de determinados grupos o reducirla en cierto grado y lograr, igualmente en cierto grado, reducir las diferencias sociales y económicas detectadas en la valoración de equidad ex-ante.

En el seguimiento de políticas destaca el análisis de incidencia presupuestaria neta (aunque también serviría como estudio ex-post). Este tipo de incidencia mide los cambios registrados en la renta real de los individuos o economías domésticas. Se incluye en esta renta la renta equivalente derivada del disfrute de los bienes y servicios públicos, producidos por la actividad presupuestaria de políticas de familia en nuestro caso. Dentro de los distintos tipos de incidencia presupuestaria, nos interesa especialmente la *Incidencia Legal o de Beneficiarios* que asigna los gastos en función de las personas en beneficio de quien se realizan. En este caso, un problema que en otras ocasiones se presenta con bastante crudeza como es el de conocer la población objetivo, es de menor importancia. Las externalidades (positivas o negativas) son mucho menores que en otras políticas y se presentan en todo caso dentro de la familia, por lo que han de tenerse en cuenta a la hora de evaluar la incidencia de políticas de tinte individual, como por ejemplo las dirigidas a las madres trabajadoras o las de servicios a la tercera edad.

Otra cuestión importante en el análisis de incidencia es la introducción de *escalas de equivalencia*, a fin de poder comparar familias con tamaño diferente. Las escalas de equivalencia son indicadores que recogen la existencia de economías de escala en la familia dependiendo de su composición, y que dan lugar a que, por ejemplo, el consumo de una familia aumente según aumentan sus miembros, pero no de forma igualmente proporcional. Se trata con ellas de conseguir hacer comparables familias con distinto tamaño. No tienen igual capacidad de gasto dos familias con igual renta cuando una está compuesta por tres miembros y otra por cinco; ni dos familias con igual renta e igual número de miembros, cuando todos son adultos o cuando hay algún niño. Las escalas de equivalencia lo que hacen es ponderar estos consumos según las edades y economías de escala para elaborar un índice por el que podríamos dividir la renta de la familia, según su tamaño y composición. A este respecto hay toda una teoría detrás que no hay que desdeñar, pero recomendamos utilizar estándares reconocidos generalmente, como las Escalas de Equivalencia de la OCDE, que facilitarían la comparación con otros estudios [7].

Conceptos relacionados, pero no equivalentes, en la equidad son la *desigualdad, progresividad, redistribución* y *pobreza*.

La *desigualdad* mide las diferencias de renta comparando los niveles en los que se coloca la población. La *progresividad* es un medio para conseguir el fin redistributivo. Una transferencia puede ser muy progresiva si incide en mayor medida sobre las rentas bajas que sobre las altas y sin embargo, puede ser escasamente redistributiva si su tipo medio es muy reducido [8]. Por tanto la *progresividad* se relaciona con la variación del tipo medio efectivo respecto a la renta y la *redistribución* se relaciona con el efecto que produce la transferencia en relación a la reducción de la desigualdad. La *pobreza* a su vez recoge aquellos que se sitúan en los niveles más bajos de renta. Para su medida es necesario primero definir que se entiende por niveles más bajos y qué punto de referencia se utiliza para considerarlos como bajos. El concepto de alto o bajo siempre lo es con referencia a una medida previamente determinada.

Las *medidas de desigualdad* parten de análisis en la participación de renta por decilas (décimas partes) o quintilas (quintas partes), comprobando qué porcentaje de población se concentra en las decilas o quintilas de renta más alta o en las más bajas.

---

[7] Un ejemplo de la importancia de las escalas de equivalencia en la medida de las políticas de familia, por ejemplo vivienda, se encuentra en Valiño, A. (2007) «Problemas de Accesibilidad a la Vivienda de las Familias en Función de su Composición y Residencia» en Familia y Economía. Estudio anual 2006. Ed Cinca. Madrid pp. 171-219.

[8] Un impuesto o una transferencia es proporcional si el tipo medio es constante para todos los tramos de renta, si el tipo medio es creciente/decreciente para todos los tramos de renta es progresivo/regresivo.

Los índices clásicos definen la *progresividad* como la diferencia entre un índice de concentración de impuestos o transferencias y un índice de desigualdad de la renta antes de la intervención pública. Los índices de *redistribución* miden la contribución de un impuesto o una transferencia a la reducción de la desigualdad mediante la diferencia entre un índice de desigualdad de la renta inicial y el mismo índice calculado después de la transferencia. Para medir el efecto redistributivo de una medida se tiene en cuenta que éste se origina como consecuencia de la interacción de tres factores: progresividad, importancia cuantitativa y el efecto reordenación [9].

El índice de desigualdad que toma en consideración todas las medidas enunciadas anteriormente es el popular *índice de Gini* (G).

$$G=1+\frac{1}{n}-\frac{2}{\mu n^2}(Y_1+2Y_{n-1}+3Y_{n-2}+\ldots\ldots+(n-1)Y_2+nY_1)$$

o también:

$$G=\frac{1}{2n^2\mu}\sum_i\sum_j|Y_i-Y_j|$$

donde:

$Y_{i,\,j}$: es la renta individual.
n: es el número de individuos
μ: es la renta media de la distribución.

Es sabido que el índice de Gini está basado en la curva de Lorenz de la distribución de la renta (L), uno de los instrumentos más conocidos y utilizados para ordenar distribuciones. Este índice cuantifica el área comprendida entre la curva de Lorenz y la bisectriz de igualdad en la distribución. Cuanto menor/mayor sea la distancia existente entre las dos curvas menor/mayor es la desigualdad que presenta la distribución analizada.

Por tanto los indicadores necesarios para evaluar la incidencia de las medidas de apoyo a la familia en términos de desigualdad son:

---

[9] El efecto reordenación rompe el vínculo entre progresividad y redistribución. La progresividad está referida a la ordenación inicial de la distribución de la población sometida a la transferencia. Sin embargo, los índices de redistribución comparan dos distribuciones la inicial y la final, posiblemente con una distinta ordenación. Esto es algo que diferencia a ambos conceptos y es necesario medir, dado que la reordenación provoca un efecto contrario a la redistribución vertical que persigue una transferencia progresiva y, por tanto, su existencia lesiona la capacidad redistributiva de la medida.

- Las *curvas de Lorenz y la curvas de concentración*.
- Las *medidas de progresividad*.
- Las *medidas de redistribución*.

Estas medidas tienen la ventaja de que pueden descomponerse; es decir, podemos conocer como contribuye cada prestación a la progresividad global o la redistribución global de un programa integral de apoyo a la familia.

Sin ánimo de ser exhaustivos, exponemos a continuación el contenido teórico en el que se apoya cada uno de los indicadores seleccionados.

- *Curvas de Lorenz y curvas de concentración*

La curva de Lorenz de una distribución de rentas es una función que nos indica la proporción de renta, respecto al total, poseída por cada porcentaje de población acumulada, una vez que hemos ordenado a los individuos de forma creciente según su nivel de renta; esto es, del más pobre al más rico. Es decir, *Ly (p)* hace referencia a la proporción de renta que posee el *p* por ciento más pobre de la población, con relación al total de renta existente en esa distribución.

La expresión analítica es la siguiente:

$$L_y(p) = L_y(j/n) = \frac{\sum_{i=1}^{j} Yi}{\sum_{i=1}^{n} Yi} = \frac{\sum_{i=1}^{j} Yi}{n\mu_y}$$

donde *m y* es la media de la distribución y, y n el tamaño de la población. *j/n* es el porcentaje de individuos que posee determinado nivel de renta.

La *curva de Lorenz*, por tanto, se construye mediante la unión entre los puntos situados en la ordenada (proporción de renta acumulada) y la accisa (propoción de población acumulada). Se trata de una curva creciente y convexa, delimitada en el intervado [0,1], de forma que necesariamente Ly(0) = 0 y Ly(1) = 1. Siempre que exista desigualdad en el reparto de la renta, los grupos más ricos poseerán un porcentaje de la renta superior a su peso demográfico y la curva de Lorenz se situará por debajo de la línea de equidistribución. Esta línea representa la distribución perfectamente igualitaria (el 10% más pobre de la población obtiene el 10% de la renta total). De esta forma, si la curva de Lorenz coincide con la línea de la

distribución igualitaria, la distribución de la renta será equitativa; si por el contrario existe distancia entre ellas, la distribución será desigual.

La *curva de Lorenz*, por tanto, no sólo constituye la forma más habitual de representar distribuciones sino que representa un instrumento sencillo para comparar el nivel de desigualdad de dos o más distribuciones. Dadas dos distribuciones x e y, decimos que la distribución x domina en sentido de Lorenz a y siempre que la curva asociada a x no se sitúe por debajo de la curva de y en ninguno de los puntos donde han sido estimadas. Esto es: Lx (p) ≥ Ly (p). El gráfico muestra que la distribución x domina a y presentado menor desigualdad.

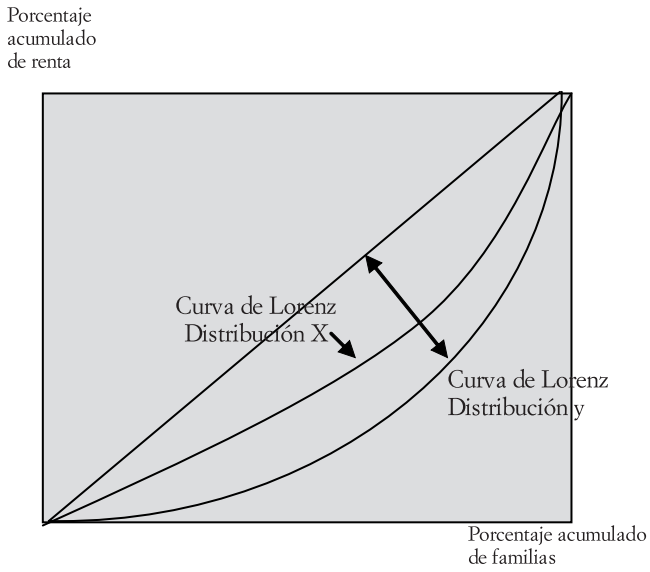


GRÁFICO 5-1

La cuantificación del área comprendida entre las curvas de Lorenz y la distribución igualitaria se calcula mediante el *índice de Gini* comprendido entre 0 y 1.

Las medidas de progresividad y de redistribución están más relacionadas con las valoraciones ex post, por lo que las incluimos en el análisis que se realiza a continuación.

Valoración ex-post o valoración final

La valoración ex-post se realiza cuando los programas llegan a su término o cuando el periodo de aplicación es suficientemente largo.

Esta evaluación toma como base las dos anteriores. Ya mencionamos que alguna de las técnicas recogidas en la evaluación ex-ante podían utilizarse en la evaluación ex-post; o al cabo de un periodo prudencial repetir la evaluación y comparar las situaciones resultantes. En los casos de evaluación de reformas estamos en un área difusa entre valoraciones ex-ante o ex-post. En general en estas evaluaciones se repiten los análisis de eficacia, de eficiencia y de equidad. En los análisis de eficiencia ahora cobran protagonismo los indicadores de resultados, y cobrando especial relevancia las comparaciones de éstos con la situación inicial evaluada antes de aplicar la política pública. Si no se hizo en su momento, sería conveniente hacerlo ahora para sopesar la importancia de los logros conseguidos.

- *Valoración de Eficiencia*

Medidas como la *valoración Coste-Beneficio* o asimiladas pueden realizarse también como valoración ex-post, especialmente cuando no se valoran alternativas. Igualmente alguna de las tipologías que se derivan de la misma, como la *Valoración Contingente*, también se aplican como valoraciones ex-post. Un ejemplo de estas evaluaciones serían, relacionadas con políticas de familia, las realizadas sobre la ayuda a domicilio en algunas Comunidades Autónomas [10].

Dentro de la tipología de estudios ex post encuadramos las técnicas de evaluación de eficiencia relativamente más complejas [11]. Dentro del análisis de eficiencia nos centramos en el análisis de *eficiencia Técnica* (que mide cantidades y se fija en los procesos productivos y en la organización de tareas).

En las líneas siguientes se recogen esquematizadas las *técnicas de evaluación de la eficiencia técnica* en función de si buscan soluciones de frontera o no. Los *análisis de frontera* buscan las mejores prácticas comparando unas unidades de producción con otras y considerando mejores aquellas que se sitúan en la frontera de producción. Las técnicas «*no de frontera*» se basan en la media de la producción.

Las aproximaciones «*no frontera*» incluyen métodos que ya hemos mencionado en la valoración ex ante y de seguimiento y, por lo tanto, no vamos a incidir más en ellas. Como novedad se buscarían funciones de producción en los modelos econométricos. Y como técnicas no mencio-

[10] Ver, por ejemplo: Martínez Argüelles, S Dávila, M. y Vicente, R (2003). Una aproximación a la evaluación económica de las políticas sociales: el caso del Servicio de Ayuda a Domicilio Revista del Ministerio de Trabajo y Asuntos Sociales Año: 2003, Número: 41, pp 89-106

[11] Aquí se va a presentar de forma muy resumida alguna de las medidas de eficiencia. Un desarrollo bastante completo se hace en Alvarez Pinilla, A (Coor.)(2001)

nadas con anterioridad habría que destacar las *series temporales*, buscando tendencias en el tiempo y efectos cíclicos; y el análisis de *Redes Neuronales*, que ayudaría a calcular funciones de producción cuando desconocemos la forma funcional (esto ocurre en casi todas las políticas públicas).

Las técnicas frontera más destacadas son el *Análisis de Envolvente de Datos* (*Data Envelopment Analysis —DEA—*) y el análisis de *Datos de Panel*. La relativa complejidad de estas medidas impide desarrollar adecuadamente sus características a efectos del análisis que estamos desarrollando. Resumimos pues la utilidad de las mismas.

El *DEA* es útil para comparar unidades que realizan servicios con distintos tipos de outputs y establecer las que realizan una mejor actividad. Es imprescindible disponer pues de varias unidades, en número relativamente elevado, realizando el mismo servicio, que además han de ser homogéneas. Todas estas condiciones son imprescindibles para poder obtener resultados significativos. Serviría para comparar los servicios a favor de las familias de una Comunidad con los servicios de igual características de otras Comunidades, o entre servicios realizados por Ayuntamientos dentro de la Comunidad o entre Ayuntamientos de distintas Comunidades, y por supuesto, entre países. El problema es que en muchos casos es difícil la existencia de la homogeneidad.

En resumen, la técnica *DEA* busca funciones frontera, es una técnica *no paramétrica* (no tiene una forma funcional previa) es una *técnica determinista* o no estocástica y permite actuar con múltiples inputs y outpus, con la condición de que sean homogéneos. Trata de encontrar el mejor conjunto de ponderaciones de inputs y outputs para la unidad analizada, con la condición de que, usando el mismo conjunto de ponderaciones, ninguna de las otras unidades de decisión obtenga un ratio de eficiencia mayor que uno.

Los métodos de análisis de *Datos de Panel* son útiles cuando las diferencias entre los individuos dependen de un elevado número de factores (diferentes de un individuo a otro) y varían con el tiempo para cada individuo. No hay que dudar mucho para asignar estas características a las familias, siendo sus diferencias condicionadas por un elevado número de factores que varían a lo largo del tiempo. Así pues, realizan observaciones de varias familias en un periodo de tiempo frente al análisis puro cross-section con observaciones de familias en un momento de tiempo y las series temporales puras con observaciones, normalmente de naturaleza agregada a lo largo del tiempo sin una dimensión longitudinal. Así pues la ventaja del análisis de datos de panel es que permite observar las mismas familias

a lo largo del tiempo. Permite seguir, pues, las reacciones que se producen en su comportamiento tras la aplicación de las políticas públicas [12].

• *Valoración de equidad*

Los análisis de equidad ex post miden la incidencia que se produce sobre la distribución de la renta de la población objetivo tras la actuación pública, midiendo los resultados de la misma sobre la equidad horizontal y/o vertical.

La intervención pública mediante impuestos o transferencias afecta a la distribución inicial de las rentas (a la que llamamos *y*), generando una distribución después (a la que llamamos *x*). Si esta intervención no hace variar el orden, la distribución inicial *y* coincide con la distribución final *x* y la posición de cada individuo coincide en las dos distribuciones. Pero es posible que se produzca un cambio en el orden de posición de los individuos tras la actuación pública, en cuyo caso la posición del individuo *i(y)* no coincide con la posición *i(x).* La alteración en el orden a que da lugar la transferencia se representa mediante una curva denominada *curva de concentración*. Por tanto, la curva de concentración de una variable cuyas observaciones están ordenadas de acuerdo a la distribución inicial indica que sus propias observaciones están desordenadas. Esto puede dar lugar a una percepción errónea sobre la efectividad de la medida en términos de reducción de la desigualdad, dado que el efecto reordenación erosiona la verdadera redistribución.

La expresión analítica de la curva de concentración es similar a la curva de Lorenz sólo que teniendo en cuenta que sus atributos están desordenados. El área entre la curva de concentración y la curva de distribución igualitaria da lugar a un *índice de concentración*, también llamado *Pseudo-Gini* comprendido igualmente entre 0 y 1.

En definitiva, lo que queremos decir, es que si un programa de impuestos o transferencias provoca un efecto reordenación en las distribuciones es necesario cuantificarlo dada la importancia que esto tiene sobre cualquier conclusión definitiva que saquemos acerca de sus efectos sobre la disminución de la desigualdad.

Las medidas de progresividad

El índice de progresividad más utilizado y conocido es el *índice de Kawani*. Basado en las curvas de Lorenz y de concentración, se define como

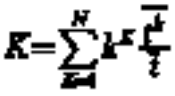
[12] Ver: Nerlove, M (2002) *Essays in Panel Data Econometrics* . Cambridge University Press.

la diferencia entre el índice de concentración de la transferencia y el indice de Gini de la distribución inicial de la renta.

$$K = Gy - CT$$

En términos gráficos supone medir el área entre la curva de concentración de la transferencia y la curva de Lorenz de la renta antes de la transferencia. Si K > 0 la transferencia es progresiva. contribuye a la redistribución de la renta y la curva de concentración se sitúa por encima de la diagonal; si K = 0 la transferencia es proporcional, no tendría ningún efecto sobre la desigualdad inicial de la renta y la curva de concentración coincide con la diagonal; y si K<0 la transferencia es regresiva, esto es, la transferencia incide de forma negativa sobre la desigualdad y la curva de concentración se sitúa por debajo de la diagonal.

El índice de Kakwani puede descomponerse para conocer cómo cada una de las transferencias que componen un programa conjunto contribuye en mayor o menor medida a la progresividad total. La expresión es la siguiente:



donde K^k es el índice de progresividad de Kawani de la transferencia k. Siendo T = T1+T2+T3+…..+Tk, y Tk >0, y donde $\bar{t}$ y $\bar{t}^k$ son los tipos medios de las transferencias totales y de la transferencia k-ésima.

Las medidas de redistribución

Para poder comparar el poder redistributivo de una transferencia recurrimos al tradicional índice de Reynols-Smolensky, que cuantifica la diferencia entre el área de las curvas de Lorenz de la distribución de la renta antes y después de la transferencia.

$$RS = G_x - G_y$$

Si las transferencias son proporcionales a la renta, la redistribución es nula y el índice de *RS*=0, en este caso, además no se produciría reordenación. Si la transferencia es progresiva el índice de *RS* >0, si es regresiva el *RS* < 0, y se podría producir reordenación en ambos casos.

Como hemos indicado anteriormente, el poder redistributivo de una determinada estructura de transferencias se origina como consecuencia de la interacción de tres factores: progresividad, cuantía de la prestación y reordenación. El índice de Reynolds-Smolensky reformulado (RS*) nos da la posibilidad de conocer de forma desagregada esa contribución a través de la expresión siguiente:

$$RS^* = \frac{\bar{t}}{1+\bar{t}} K + R$$

siendo R el término reordenación de Atkinson-Plonick que cuantifica la distancia entre la curvas de concentración y la curva de Lorenz de la renta después de la transferencia. Podemos incluso descomponer el índice en redistribuciones parciales. Es decir, podremos conocer cómo las diferentes transferencias que componen un programa contribuyen a la redistribución global del mismo, mediante la siguiente expresión:

$$RS^* = \sum_{k=1}^{N} K^k \frac{\bar{t}^k}{1+\bar{t}} + R$$

Si R = 0; esto es, la reordenación es nula, entonces:

$$RS^* = \sum_{k=1}^{N} K^k \frac{\bar{t}^k}{1+\bar{t}}$$

En el caso de que se produzcan reordenaciones la posibilidad de descomposición no está muy clara, dado que la suma de las reordenaciones de cada una de las transferencias no tiene por que sumar el total. Pueden producirse compensaciones entre unas y otras que anulen los efectos parciales de cada transferencia (Pazos y Salas, 1995).

En las medidas de equidad de las reformas o cambios de política también resultan de interés alguna de las medidas que hemos presentado en el análisis de eficiencia. Por ejemplo el análisis de panel de datos dinámico en el que se observa como queda la población objetivo tras la realización del programa y comparándolo con la situación en la que estaban antes de su aplicación. Se representan en una matriz, de forma que los que se sitúan por encima de la diagonal ven mejorada su situación tras las reformas, lo que están sobre la diagonal permanecen igual y los que están por debajo empeoran. Viendo a que niveles de renta pertenecen antes y después, podemos evaluar la actuación redistributiva o no del proyecto analizado.

5.3. CARACTERÍSTICAS Y TIPOS DE INDICADORES. AGRUPACIÓN DE LOS INDICADORES POR CRITERIOS Y OBJETIVOS

Según la Unión Europea, «un indicador puede definirse como una característica o atributo que puede medirse con el fin de evaluar el progreso de un programa hacia la consecución de sus objetivos».

Como antes indicamos, disponer de un conjunto adecuado de indicadores es un elemento básico de la evaluación de políticas públicas, en sí mismos y como punto de partida para la aplicación de otras técnicas.

Su simplicidad permite el manejo y conocimiento de los no iniciados en técnicas más complejas como las recogidas en el análisis anterior. Aunque como en otros muchos campos, la simplicidad impida conocimientos más profundos.

En principio, los indicadores son siempre medidas relativas. La relación a la que se refieren se expresa en cociente y prácticamente siempre en porcentaje.

El análisis que hemos realizado en las páginas anteriores ponía de relieve las relaciones de los indicadores con la evaluación ex ante, durante y ex post a la aplicación de las políticas públicas, y cómo se relacionan especialmente con la valoración de cumplimiento de objetivos o ejecución. En este apartado vamos a desarrollar los indicadores necesarios para realizar el seguimiento de las políticas de familia, planteando los posibles objetivos a los que estas pudieran atender. Estos objetivos dependen a su vez de los problemas a los que se enfrentan las familias, para cuya detección deberán plantearse indicadores de necesidad. Los programas a aplicar dependen a su vez de los objetivos a resolver y los indicadores de ejecución pondrán en relación la población objetivo con los objetivos planteados. A pasar de que se presentan un gran número de indicadores, la evaluación de una política concreta podría exigir que fueran otros, o sólo alguno de los que aquí se presentan.

Un paso inicial es el conocimiento de los problemas que se van a intentar resolver por las actuaciones públicas. Es lo que nosotros hemos denominado *indicadores de necesidad* y la UE denomina *indicadores de contexto*. En estos indicadores deben de recogerse las características sociales, demográficas y económicas de las familias.

Diagnóstico de la situación. Problemas a intentar resolver. Indicadores de contexto (necesidades)

El interés por las políticas de familia surge de una demanda social y por la necesidad de potenciarla como medio para resolver otros problemas

para el erario público. En la tabla siguiente se recoge, a modo de ejemplo, un resumen de las características demográficas, sociales y económicas que podrían dar lugar a generación de necesidades, indicadores de contexto para evaluarlas y posibles diagnósticos o necesidades a satisfacer.

TABLA 5-1

Características	Indicadores de necesidad	Diagnóstico - Necesidades
Demográficas: Bajas tasas de natalidad. Alargamiento de la esperanza de vida. Bajas tasas de mortalidad.	• Tasa de natalidad. • Esperanza de vida a los 70. • Morbilidad en mayores de 70 años. • Tasa de mortalidad. • Población por edad (mayores de 65, de 70, 75, 80). y sexo (hombres, mujeres) respecto de la población total • Tasa de cobertura: Población mayor de 65 años. respecto de la población entre 16 y 65 años. • Población menor de 16 años por sexo respecto de la población total. • Tasa de envejecimiento: población menor de 16 años respecto de la población mayor de 65 años.	**Envejecimiento** o no de la población: Necesidad de cuidados –aumento de la dependencia de ancianos–: (Tareas domésticas, servicios sanitarios, servicios sociales). Problemas en la cobertura de las pensiones.
Sociales: Matrimonios Divorcios Maternidad- fecundidad. Edades a las que se accede a la primera maternidad. Madres por primera vez mayores de 39 años. Madres quinceañeras.	• Porcentaje de matrimonios o parejas (en relación a las personas en edad fértil). • Porcentaje de divorcios. • Porcentaje de nuevas nupcias por edad y sexo. • Tasa de fecundidad. • Tasa de fecundidad por tramos de edad. • Edad media a las que se accede a la primera maternidad. • Porcentaje de mujeres que acceden por primera vez. a la maternidad después de los 39 años. • Porcentaje de madres menores de 16 años. • Tamaño medio de las familias. • Porcentajes de familia sin hijos, con 1 solo hijo, con 2 hijos, con 3 o más hijos. • Tasa de natalidad de nacionales y extranjeros.	**Tasas de natalidad.** Necesidad de elevar la tasa de natalidad.
Sociales: **Conflictos familiares** Malos tratos a mujeres, niños, discapacitados, ancianos, de los hijos a los padres.	• Porcentaje de mujeres maltratadas, por niveles educativos y renta. • Porcentaje de menores de 18 años maltratados por sexo, por niveles educativos de los padres y renta. • Porcentaje de mayores de 65 maltratados por sexo, niveles educativos (propios y familiares) y renta propia y familiar. • Porcentaje de discapacitados maltratados por sexo, edad, nivel educativo familiar y renta.	**Reducir / eliminar las tasas de malos tratos:** Introducir sistemas para facilitar la relación familiar, garantizando la integridad de sus miembros.

TABLA 5-1 *(Continuación)*

Características	Indicadores de necesidad	Diagnóstico - Necesidades
Sociales: **Conflictos sociales (*)** Huelgas. Suicidios. Crímenes. Crímenes juveniles. Población reclusa. Población reclusa juvenil. Drogadicción, muertes por drogadicción. Grado de solidaridad Civismo.	• Días no trabajados por cada 1000 empleados asalariados. • Tasa de suicidios por sexo , edad, renta y tipo de familia por cada 1000 personas. • Víctimas de crímenes y de violencia una vez o más. • Crímenes juveniles por sexo, edad, renta y tipo de familia. • Adultos convictos en prisiones por sexo, edad, estado civil y número de hijos por edades. • Tasa de drogadicción por edad, renta y tipo de familia. • Muertes relacionadas con la drogadicción por 1000 personas. • Uso de cannabis y las anfetaminas por menores de 18 años, niveles de renta y tipo de familia.	**Reducir el nivel de conflicto social:** Relacionado con la reducción de la pobreza, de los conflictos familiares.
Económicos		
Laborales: Incorporación de la mujer al mercado laboral Empleo/ desempleo femenino. Tipos de trabajo Distribución de tiempo/ trabajo familia.	• Diferencias en remuneraciones laborales por sexo. • Por sexo, edad, educación, estado civil, número de hijos, hijos menores de 1, de 3 y de 6 años, nivel de renta, con mayores dependientes según edades y discapacitados: – Tasa de actividad. – Tasa de paro. – Frecuencia del paro de larga duración. – Frecuencia del empleo total a tiempo parcial. – Frecuencia del empleo temporal. – Sector de empleo. • Estructura de la actividad laboral por tipos de familia. (Parejas con hijos menores de 6 años y mayores –por edades–). • Tasas de abandono de trabajo por motivos familiares, laborales, personales. • Absentismo laboral por maternidad, cuidados, enfermedad, y razones personales. • Distribución de tiempo tareas domésticas, cuidados de dependientes, ocio y trabajo. • Por sexo, edad a partir de años 65, renta: mayores solos, en residencias y en familia respecto a la población total mayor. • Tasas de escolaridad por grupos de edad (menores de 1, 2,3 y entre 3 y 6).	**Conciliar vida familiar y laboral:** Cuidados a niños. Cuidados a ancianos. Cuidados a discapacitados. Igualdad de sexos.

TABLA 5-1 *(Continuación)*

Características	Indicadores de necesidad	Diagnóstico - Necesidades
Bienestar familiar: Cargas familiares. Vivienda. Uso de nuevas tecnologías. Participación social y cultural. Familias en situación de conflicto y ruptura.	• Desigualdad en la renta: Ratio de participación en quintilas (ratio del total de renta recibida por el 20% de la población del país con la mayor renta (quintila superior) en relación con la recibida por el 20% de la población con menor renta (menor quintila). • Coeficiente de Gini. • Renta media de las familias por número de miembros, edades, y tipos de dependientes. • Tipos de acceso a la vivienda (propiedad, alquiler) por tipo de familia y nivel de renta. • Tamaño de la vivienda y características por tamaño de familia (m^2, por miembro de la familia y deficiencias estructurales). • Renta media de las familias con algún miembro discapacitado. • Número de ordenadores por familia. • Tasa de acceso a Internet. • Número de teléfonos móviles por familia. • Frecuencia de accesos a teatro, cine y espectáculos, tiempo dedicado a la lectura. • Tasa de separaciones.	**Disminuir las diferencias en las rentas** Apoyo a: · Acceso a la vivienda. · Acceso a la vivienda a familias numerosas. · Acceso a nuevas tecnologías. · Acceso a la cultura. · Reducción de conflictos familiares.
Rentas bajas: **Pobreza** **Familias con riesgo de exclusión social** Familias numerosas. Familias monoparentales. Familias en paro. Familias en paro de larga duración. Discapacitados. Los «sin techo». Exclusión por falta de acceso a nuevas tecnologías. Familias inmigrantes. Familias gitanas.	INDICADORES MONETARIOS • Por tamaño familiar, por tipo de paro, por discapacidad de los miembros, por raza, por grupos étnicos, por nacionalidad, por tipo de vivienda, por acceso a tecnologías: – Familias por debajo del 15% de la renta media (pobreza extrema). – Familias entre el 16% y 25% de la renta media (pobreza grave). – Familias entre el 26 % y 35% (pobreza moderada). – Familias entre el 36% y el 50% de la renta media (precariedad social). – Familias por debajo del 50% de la renta media (información sobre varios años a tras indicaría la permanencia en estos niveles). INDICADORES NO MONETARIOS • Por tamaño familiar, por tipo de paro, por discapacidad de los miembros, por raza, por grupos étnicos, por nacionalidad, por tipo de vivienda, por acceso a tecnologías y tipo de pobreza: • Porcentajes de familias en las que nadie trabaja. • Porcentaje de familias con niños sin escolarizar. • Porcentajes de familias con viviendas con algún tipo de precariedad (por tipo de tenencia de vivienda). • Porcentaje de familias sin vivienda (sería necesario disponer de estadísticas y estudios sobre los «sin techo» que suelen ser personas solas y en desarraigo familiar, por lo que colateralmente entran en el estudio).	**Eliminar la trampa de la pobreza:** Lograr que el mayor porcentaje de familias posible alcancen niveles de renta superiores. Mejorar las tasas de empleo en estos grupos.

(*) Características e indicadores extraídos de OCDE

En la siguiente tabla se recogen los posibles objetivos de las políticas de familia, con la precisión de que su ordenación por importancia o preferencia en la atención, y grado en el que han de obtenerse, están en relación con las necesidades descubiertas a través de los indicadores antes enumerados y las decisiones políticas del gobierno existente. La tabla 5-1 anterior recogía una columna en la que se enumeraban posibles necesidades resultantes, completamos esta columna con la siguiente tabla en la que recogemos posibles objetivos a conseguir. Se clasifican en objetivos finales, intermedios y operacionales, siguiendo las clasificaciones comentadas anteriormente.

Los objetivos operacionales se refieren a medidas concretas (por ejemplo determinado número de guarderías, o determinado número de plazas para ancianos, etc.). Han de ser necesariamente explicitados por el político, así que nos limitamos a plantear el grupo de políticas en las que se incluirían, cuyas posibilidades de desarrollo realizaremos en la tabla 5-3.

TABLA 5-2

OBJETIVOS			
Finales	**Intermedios**		**Operacionales**
Incrementar la calidad de vida de las familias	Conciliación de la vida familiar y laboral.	Ayuda a la tercera edad como dependientes. Apoyo en los: Cuidados a ancianos. Cuidados a niños. Cuidados a discapacitados. Igualdad de sexos.	• Políticas y medidas de Conciliación.
	Facilitar el acceso a la vivienda.	Apoyo a: · Acceso a la vivienda. · Acceso a la vivienda a familias numerosas.	• Políticas de vivienda.
	Mejorar los niveles de renta.	Disminuir las diferencias en las rentas.	• Políticas de rentas.
		· Acceso a nuevas tecnologías. · Acceso a la cultura. · Reducción de conflictos familiares.	• Políticas culturales.
			En general apoyo a las familias en políticas paralelas que facilítenle acceso a los servicios (educación, sanidad y transporte).

TABLA 5-2 *(Continuación)*

OBJETIVOS			
Finales	**Intermedios**		**Operacionales**
Fomentar la solidaridad intergeneracional.	**Tasas de natalidad.**	Necesidad de elevar la tasa de natalidad	• Políticas de conciliación. • Políticas estrictas de carácter natalicio.
Apoyar a la familia como garante de la cohesión social.	**Reducir / eliminar las tasas de malos tratos.**	Introducir sistemas para facilitar la relación familiar, garantizando la integridad de sus miembros.	• Políticas de rentas. • Servicios sociales específicos.
	Reducir el nivel de conflicto social.	• Relacionado con la reducción de la pobreza, de los conflictos familiares. • Apoyar la solución pacífica de conflictos. • Entorno jurídico y económico que favorezca y premie a las familias españolas que asumen responsablemente la función de solidaridad. • Familias en cuyo seno se produce violencia.	• Políticas de integración. • Políticas de rentas. • Servicios sociales específicos. • Regulación.
Apoyo a las familias en situación de riesgo social y otras situaciones especiales.	**Eliminar la trampa de la pobreza:** Lograr que el mayor porcentaje de familias posible alcancen niveles de renta superiores. Mejorar las tasas de empleo en estos grupos.	Apoyo en los niveles más bajos de renta a: Familias en situación de desempleo Inserción sociolaboral. Familias monoparentales, familias en situación de conflicto o ruptura. Familias con discapacitados. Familias en cuyo seno se produce violencia: – violencia contra la mujer – violencia contra menores – maltrato a ancianos violencia de los hijos hacia sus padres	• Políticas de renta. • En general (siempre ligadas a rentas) políticas de atención específica a: Familias monoparentales. Padres desempleados. Familias numerosas. Familias inmigrantes y grupos étnicos marginados.

A continuación en la tabla 5-3 recogemos un estudio de las posibles políticas de familia a instrumentar para alcanzar los objetivos propuestos. De nuevo escoger el tipo de política y el grado en que se aplica será una cuestión a decidir por el político.

TABLA 5-3

Política	INSTRUMENTOS			Población objeto
	Tipo	Medida	Carácter	
Políticas y medidas de Conciliación.	Regulación	Disposiciones sobre permisos parentales	Específico	Padres con hijos menores de 3 años
		Disposiciones sobre permisos de cuidados	Universal	familias
	Beneficios fiscales	Ayudas en el IRPF a la familia (deducciones en general que gradúan la capacidad de pago según miembros)	Universal (graduación en función de la dimensión familiar)	familias
		Ayudas en el IRPF a madres trabajadoras por reincorporarse inmediatamente al trabajo (en realidad son ayudas laborales, no familiares)	Específico	Madres trabajando con hijos menores de 3 años
		Ayudas a familias con hijos menores de 3 años. (desgravaciones de gastos por guardería o cuidadores)	Específico	Familias con hijos menores de 3 años
		Ayudas a familias con mayores a su cuidado (ayudas directas generales o desgravaciones por cuidados especiales)	Específico	Familias con mayores a cargo
		Ayudas a familias con discapacitados a cargo	Específico	Famillias con discapacitados
	Servicios sociales personales	Guarderías Asilos Centros de día para mayores Servicios de teleasistencia Servicios de ayuda en el hogar o ayuda a domicilio Servicios sanitarios para situaciones especiales	Universales	familias
	Gastos directos	Subvenciones para cuidados de: Niños Ancianos Discapacitados	Universales	familias
	Seguridad Social	Ayudas en la cobertura de cuotas empresariales de la SS durante los periodos de baja por maternidad y permisos parentales	Específico	Padres en situación de permiso

TABLA 5-3 *(Continuación)*

Política	INSTRUMENTOS			Población objeto
	Tipo	**Medida**	**Carácter**	
Políticas de vivienda	Regulación	Disposiciones generales sobre características y requisitos mínimos	Universales	familias
	Fiscales	En el IRPF: Descuentos por compra de vivienda habitual Descuentos por alquiler de vivienda Descuentos en gastos de prestamos hipotecarios En el IBI: descuentos en la base o en la cuota Descuentos en tasas municipales En IPN: descuentos en la base ITP Y AJD: descuentos en las trasmisiones	Universales (graduación en función de la dimensión familiar).	Familias
		I. Sucesiones: eximir a menores en sucesiones directas		
	Gastos directos	Tipos de interés hipotecario subvencionado VPO (teniendo en cuenta la necesidad de incluir viviendas para familias numerosas)	Universales (graduación en función de la dimensión familiar)	Familias
Políticas de rentas	Fiscales	Ayudas en el IRPF a la familia (deducciones en general que gradúan la capacidad de pago según miembros) Exención en impuesto de circulación de vehículos a familias numerosas Exención IVA en vehículos de gran dimensión para familias numerosas	Universal (graduación en función de la dimensión familiar)	familias
	Gastos directos	Subvenciones a familias en general: • Por matrimonio • Por hijos • Por otros dependientes	Universal (graduación en función de la dimensión familiar)	familias
		Subvenciones especiales para familias monoparentales	Específicas	Familias monoparentales
		Rentas mínimas de subsistencia	Con límite de rentas (graduación en función de la dimensión familiar)	Familias por debajo del límite de renta

TABLA 5-3 *(Continuación)*

Política	INSTRUMENTOS			Población objeto
	Tipo	Medida	Carácter	
	Seguridad Social	Suplementos o modificaciones de los beneficios en caso de: Enfermedad Accidentes o enfermedades laborales Invalidez Retiro Desempleo Beneficios para viudas/ viudos Beneficios para huérfanos	Universales (graduación en función de la dimensión familiar)	familia
Políticas culturales	Fiscales	IRPF: Deducciones o desgravaciones por estudio de hijos universitarios o de Formación Profesional	Específicas (graduación en función de la dimensión familiar), con límites de renta	Familias con hijos en edad de realizar estudios superiores
En general apoyo a las familias en políticas paralelas que faciliten el acceso a los servicios (educación, sanidad y transporte)	Gastos directos	Subvenciones educativas Becas por estudios Préstamos para estudios universitarios	Específicas (graduación en función de la dimensión familiar), con límites de renta	Familias con hijos en edad de realizar estudios superiores
		Abonos de transporte	Específicos familias numerosas	
		Subvenciones a las ediciones de libros de literatura para jóvenes y niños (las recibirían las editoriales a fin de reducir el precio)	Universales	familias
	Servicios sociales	Actividades culturales municipales (teatros, exposiciones, lectura de cuentos, formativas, etc.)	Universales (dimensión familiar)	Familias
		Servicios de acceso a internet municipales Ayudas a colegios públicos para incorporar acceso a líneas ADSL con fines educativos		

TABLA 5-3 *(Continuación)*

Política	INSTRUMENTOS			Población objeto
	Tipo	Medida	Carácter	
Políticas estrictas de carácter natalicio. (Ayudan las políticas de conciliación)	Regulación	Disposiciones sobre la paternidad Regulación de derechos de los niños Regulaciones sobre la adopción	Universal	Familias
	Fiscales	IRPF: Descuentos especiales por nacimiento de hijos.	Universales (graduación en función de la dimensión familiar)	Familias
	Gastos directos	Subvenciones por nacimiento de hijos (cantidades que pueden mantenerse periódicamente durante los tres primeros años de vida del niño).	Universales (graduación en función de la dimensión familiar)	Familias
	Servicios sociales	Servicios asistenciales a las madres embarazadas (sanitarios y formativos) Servicios asistenciales de consejo sobre maternidad o aborto. Servicios de acogida de niños abandonados.	Universales	Familias
Para la cohesión social Servicios sociales específicos (Ayudan las políticas de rentas, las de bienestar familiar y conciliación)	Regulación	Regulación sobre malos tratos. Regulación sobre estancia en prisiones madres con hijos menores. Regulación sobre separaciones y divorcios.	Universales	Familias
	Servicios sociales	Servicios asistenciales a drogodependientes. Servicios asistenciales a mujeres maltratadas. Servicios de acogida a niños con malos trato. Servicios de atención a niños y mayores dependientes en casos de separación o divorcio. Servicios de consejo en casos de conflicto familias. Servicios de reintegración.	Universales	Familias

TABLA 5-3 *(Continuación)*

Política	INSTRUMENTOS			Población objeto
	Tipo	Medida	Carácter	
Políticas para evitar la exclusión social y ayudas a la pobreza. En general (siempre ligadas a rentas) políticas de atención específica a: Familias monoparentales Padres desempleados Familias numerosas Familias inmigrantes y grupos étnicos marginados	Gastos directos	Rentas mínimas	Límite de rentas	Familias por debajo del límite de renta establecido
		Ayudas especiales a familias numerosas	Específico familias numerosas con límite de rentas	Familias numerosas por debajo del límite de renta establecido
	Servicios sociales	Apoyo especial a familias en grupos de riesgo (numerosas, parados, inmigrantes, grupos étnicos de riesgo, con discapacitados) con acceso gratuito a: Guarderías. Centros asistenciales. Centros de consejo e información. Centros de acogida. Programas de educación familiar. Servicios de teleasistencia y Ayuda a domicilio a mayores. Ayudas a ONGs que trabajen con los grupos de exclusión o riesgo (directas a las instituciones).	Límite de rentas	Familias dentro del límite de rentas.
	Políticas de vivienda	Ayudas al acceso a la vivienda específicas para rentas bajas VPO.	Límite de rentas	Familias dentro del límite de rentas

La tabla número 5-4 siguiente recoge los elementos o datos que serían imprescindibles para la valoración con indicares de seguimiento y ex-post. Su contenido estará en relación con los objetivos y políticas establecidos en las tablas anteriores que también condicionan los indicadores que han de recogerse en la tabla 5-5.

TABLA 5-4

ELEMENTOS DE VALORACIÓN				
Población objetivo	**Beneficiarios**	**Costes totales**	**Número de Actuaciones**	**Presupuesto disponible**
Aquel grupo al que se pretende atender, que a su vez depende de los objetivos inicialmente establecidos: Cuando la política es de carácter universal, la población objetivo son todas las familias Cuando la política tiene carácter específico (p. e.: familias numerosas) la población objetivo es aquella que se especifica Cuando la política tiene carácter limitado en rentas, la población objetivo será la que se encuentra dentro del grupo de renta fijado en las condiciones.	En principio deberían coincidir con la población objetivo, pero no siempre es así. Los motivos pueden ser varios: Un mal diseño de la política que quería p.e. ser universal pero el sistema de aplicación lo impide. Esto puede ocurrir cuando utilizamos un objetivo o un instrumento intermedio para alcanzar un objetivo final. También puede producirse por falta de información. Las familias en cuestión desconoce que disponen de ese beneficio.	Recoge los costes del proyecto. Los costes que han de incluirse son: 1. Directos de la aplicación del mismo: gastos fiscales, o gastos en transferencias, o de costes de personal y compras de bienes y servicios, elementos de mobiliario... etc. 2. Costes imputados derivados de la gestión del servicio, amortización de inmuebles, etc.	Información necesaria para evaluar la productividad. Deberá incorporar información sobre lo que produce la unidad evaluada (p.e. si es un servicio de teleasistencia: número de atenciones telefónicas realizadas al día o la hora)	Cantidades asignadas a las unidades operativas en presupuesto anual y las previstas en el periodo de vida del proyecto (si es limitado) o en el plan global de actuación.

La siguiente tabla recoge los indicadores que serían necesarios para la evaluación de seguimiento y final. Así pues, el contenido de la tabla depende de los objetivos establecidos, de las políticas aplicadas en concreto, y en consecuencia de la población objetivo. Utilizaría la información recogida en la tabla 5-4 y la combinaría con los indicadores de necesidad establecidos en a tabla 5-1. De hecho la mayoría de los indicadores de eficacia que se recogerían en el cuadro 5 serían diferencias en el tiempo de estos últimos indicadores antes de aplicar las políticas y después de aplicarlas. En los casos en los que la política ya estaba funcionando con anterioridad, se actúa comparando respecto a la situación hipotética de que no se estuviera aplicando la política (por ejemplo comparando nivel de renta antes y después de una subvención). A título de ejemplo se presentan algunos posibles indicadores para el caso de ayudas a la conciliación de vida familiar y laboral y para el caso de ayudas de renta a familias numerosas.

TABLA 5-5

INDICADORES						
Necesidad	**Cobertura**	**Costes**		**Volumen de actividad**	**Eficacia**	
Población Objetivo / Total	**Beneficiarios / Población objetivo**	**Medios**	**Por beneficiario**	**Productividad**	**Ind. directos**	**Ind. Indirectos**
Familias con hijos < 3 años / total familias con hijos	Familias con hijos < 3 años que reciben ayuda / familias con hijos < 3 años	CT / n.º guarderías	CT / niño CT / familia	N.º plazas / n.º niños < 3 años	N.º solicitudes sobre plazas ofertadas	Incremento de la natalidad. Incremento de la tasa de empleo de las mujeres
Familias numerosas / total familias	Familias numerosas que reciben ayuda de gasto directo / total familias numerosas	CT / n.º ayudas	CT / niños	Tiempo medio de tramitación de la ayuda	Mejora de nivel de renta de las familias numerosas	Incremento de la natalidad

5.4. BASES DE DATOS

Se recoge aquí, a título informativo, algunas de las fuentes de datos que pueden ser útiles para el análisis de las políticas de familias. Es necesario especificar que no están, ni mucho menos, todas las fuentes de datos. A nivel internacional, destacan los datos de la OCDE. Esta organización se ocupa sobre todo, en lo referente a familia, de estudios sobre conciliación de vida familiar y laboral y de estudios sobre dependencia de mayores. La Unión Europea también presenta datos de este orden. A través del Eurostat se puede obtener información de los hogares, ya sea por medio de las encuestas de presupuestos familiares (Panel de Hogares de la Unión Europea). El Fondo Monetario Internacional también ofrece algunos datos útiles, o la Organización Internacional del Trabajo.

Pero la fuente más importante de datos para el estudio de las familias va a ser el Instituto Nacional de Estadística del país cuya política se esté evaluando, sobre todo a través de las encuestas que realice:

- la encuesta de presupuestos familiares,
- la encuesta de usos del tiempo,
- la encuesta de población activa,
- la encuesta sobre discapacidades, deficiencias y estado de salud,

- la encuesta de morbilidad hospitalaria,
- la encuesta sobre el tiempo de trabajo,
- la encuesta del coste de mano de obra,
- la encuesta sobre salarios en la industria y servicios,
- la encuesta de fecundidad,
- la encuesta sobre equipamiento y uso de tecnologías de información y comunicación en los hogares.

A esto se añaden estadísticas capaces de proporcionar alguna variable que resulte necesaria para la evaluación como:

- Estadísticas sobre la producción editorial de libros.
- Estadísticas de la enseñanza.
- Estadísticas sobre hipotecas.
- Estadísticas sobre extranjeros residentes.

El Ministerio de Trabajo y Asuntos Sociales del país correspondiente puede ser una buena fuente de información referente a temas laborales, permisos parentales, guarderías laborales o servicios a la tercera edad

Los problemas que pueden aparecer es la falta de datos a nivel municipal. Para algunos de los tipos de evaluación que hemos mencionado a lo largo del estudio son necesarias encuestas directas entre los usuarios efectivos o potenciales de los servicios. En otros casos el problema es la especificación de datos para la familia, ya que aparecen los datos individualizados, o en el mejor de los casos se refieren a hogares, que no necesariamente son el concepto de familia (por ejemplo el caso de dos amigos o estudiantes que comparten domicilio, o la utilización por varias familias —p. e. inmigrantes— de la misma vivienda). No obstante la proporción de estos casos en las encuestas de hogares suele ser muy pequeña por lo que se suelen asimilar ambos conceptos.

BIBLIOGRAFÍA

ALVAREZ PINILLA, A (2001) *La Medición de la Eficiencia y la Productividad.* Coordinador. Ed. Pirámide.

ATKINSON, A.B. (1970). «On the measurement of inequality», *Journal of Economic Theory,* 2, pág. 244-263.

BADENES, N.; LÓPEZ LABORDA, J.; ONRUBIA FERNÁNDEZ, J. (2000) «Efectos redistributivos y sobre el bienestar social del tratamiento de las cargas familiares en el nuevo IRPF». *Documentos de Trabajo.* Fundación de las Cajas de Ahorros Confederadas para la Investigación Económica y Social n.º 167/2001.

CUADRAS, C. (2004) *Análisis multivariante,* Universidad de Barcelona. ecology research, Springer.

BARDACH, E. (2000): *A Paractical Guide for Policy Análisis, The Eightfoldl Path to More Effective Poblem Solving.* New York, NH: Chatham House Publishers.

JOHNSON, R.A. Y WICHERN, W.A. (2002) *Applied multivariate statistical analysis,* Prentice Hall.

KAKWANI, N.C. (1977). «Application of Lorenz Curves in Economic Analysis», *Econometrica,* 45, pág. 719-727.

KALISCH, D.W.; AMAN, T. Y BUCHELE, L.A. (1998) «Social and Health Policies in OECD countries: a survey of current programmes and recent developments. Labour market and social policy». *Occasional papers* n.º 33.

LAMBERT, P.J. (1993). *The Distribution and Redistribution of Income,* 2ª Ed. Manchestet University Press, Manchester. (v.c. traducida por el IEF, 1996).

LÓPEZ LÓPEZ, M.ª T. (1997) *La protección social a la familia en España y en los demás estados de la Unión Europea.* Fundación BBV. Serie Economía Pública. Bilbao.

LOPEZ, M.T., VALIÑO A. (2004) «Políticas Públicas de Conciliación en la Comunidad de Madrid. Un análisis de las variables del PHOGUE». En *Proyecto Equal: Conciliación Una Condición para la Igualdad. Investigaciones promovidas por el Consejo de la Mujer de la Comunidad de Madrid* pp. 259-332. ISBN: 84-921275-6-2.

LOPEZ, M.T., VALIÑO A. (2004) *Políticas Públicas de conciliación de la vida familiar y laboral en la Unión Europea. Análisis de sus efectos económicos.* Ed. CES.

LOPEZ, M.T., VALIÑO A. (2004) «Familia y Conciliación de la vida familiar y laboral». En el *Informe de la Fundación Acción Familiar. Los últimos 20 años de políticas de ayuda a la familia en España.* En prensa.

MARTÍNEZ ARGÜELLES, S DÁVILA, M. Y VICENTE, R (2003) «Una aproximación a la evaluación económica de las políticas sociales. EL caso del Servicio de Ayuda a Domicilio». *Revista del Ministerio de Trabajo y Asuntos Sociales* n.º 42, p 89 a.

MUNDA, G. (2003) *Multicriteria Assessment.*
http://www.ecologicaleconomics.org/publica/encyc_entries/Mlticritassess.pdf

NERLOVE, M (2002) *Essays in Panel Data Econometrics*. Cambridge University Press.

OBSERVATOIRE EUROPEEN DES POLITIQUES FAMILIALES NATIONALES (1992). *Les politiques familiales nationales des etats membres de la Communauté Europeenne en 1991.* DGV. Bruxelles.

PABLOS, L. VALIÑO , A (2000) *Economía del Gasto Público: control y evaluación.* Ed. Cívitas. Colección Economía.

PAPACOSTANTINOU, G Y POLT, W, (1997) Policy Evaluation in innovation and technology: an overview. Chapter 1, OCDE Conference on Policy Evaluation in Innovation and Technology, 26-27 junio.

PAZOS, M. Y SALAS, R. (1995). «Descomposición de la progresividad y redistribución de las transferencias públicas». *Papeles de Trabajo,* 14.

PEÑA, D. (2002) *Análisis de datos multivariantes,* McGraw-Hill.

ROY B. (1985) - *Méthodologie multicritère d'aide à la décision,* Economica, Paris.

ROY B. (1996) - *Multicriteria methodology for decision analysis,* Kluwer, Dordrecht.

SAATY T.L. (1980) - *The analytic hierarchy process,* McGraw Hill, New York.

SALAS, R. (2001)**.** «La medición de la desigualdad económica». *Papeles de Trabajo,* 14. Instituto de Estudios Fiscales.

VALIÑO A. Y LÓPEZ T. (2006) "Factores explicativos de las ayudas directas a las familias por Comunidades Autónomas". *Documentos de Trabajo de la Facultad de Ciencias Económicas y Empresariales,* n.º 5, 2006. http://www.ucm.es/BUCM/cee/050501.htm#13.

VALIÑO, A. (2007) «Problemas de Accesibilidad a la Vivienda de las Familias en Función de su Composición y Residencia» en *Familia y Economía.* Estudio anual 2006. Ed Cinca. Madrid pp. 171-219. ISBN: 978-84-935104-7-3.

VINCKE PH. (1992) - *Multicriteria decision* aid, Wiley, New York.

YITZHAKI, S. (1983). «On an Extension of the Gini Inequality Indix», *International Economic Review,* 24, pág. 617-628.

ZELENY M. (1982) - *Multiple criteria decision making,* McGraw Hill, New York.

CAPÍTULO VI
LA DISTRIBUCIÓN Y DESIGUALDAD DE LA RENTA

NURIA BADENES PLÁ

6.1. INTRODUCCIÓN

El presente capítulo aborda el estudio de la distribución de la renta y de la desigualdad que se deriva de dicha distribución. La preocupación por la desigualdad en el campo de investigadores y académicos no es sino reflejo de los intereses de la sociedad: a los ciudadanos les preocupa la desigualdad, ya sea porque son los individuos peor situados, bien porque temen ocupar los lugares más atrasados en una distribución, bien por solidaridad hacia los peor situados o simplemente por la incomodidad, la pesadumbre de conciencia o posibilidad de conflictos ligada a la desigualdad. Si los ciudadanos incluyen las cuestiones distributivas entre sus intereses, los políticos y decisores sociales deberán incluirlas también en sus agendas, ya porque desean satisfacer a la población que representan o sea porque pretenden maximizar la probabilidad de ser elegidos o reelegidos.

La investigación relacionada con cuestiones distributivas ha sido muy fructífera en las últimas décadas, especialmente a medida que se ha ido disponiendo de mayor información estadística, de datos desagregados y al mismo tiempo, se ha avanzado en el uso de técnicas analíticas más modernas. Pero el estudio de la distribución y desigualdad de la renta no constituyen —como señala Atkinson— una materia prioritaria en el análisis económico, más centrado en cuestiones relativas a productividad y eficiencia. Ello explica en parte que el cuerpo teórico y el desarrollo de herramientas de análisis distributivo haya sufrido cierto retraso con respecto a otras cuestiones económicas.

El estudio de la distribución de la renta es necesario si se quieren abordar cuestiones básicas en el ámbito de la Economía del Bienestar. Por ejemplo la medición de la desigualdad, la comparación de situaciones previas y posteriores a una medida de actuación del gobierno (ya sea cobro de impuestos o reparto de transferencias), el análisis de la incidencia (o estudio de afectados) por determinada política, o simplemente la descripción de un colectivo para su comparación con otros similares o a lo largo del tiempo o con otros colectivos.

La elección de la renta como variable descriptiva del bienestar de los individuos no es casual. En definitiva, nos interesa calcular lo bien que se en-

cuentran los individuos o las familias, y dado que la utilidad y el bienestar no son fácilmente medibles (aunque sí comparables) [1], podríamos utilizar variables que hagan más felices al individuo, como pueden ser la renta, la riqueza o el consumo. Hay que recalcar que nuestro punto de vista es estrictamente económico, por lo que no nos interesan situaciones en la que un rico se siente desgraciado a pesar de poseer mucha renta y poder acceder a mucho consumo o un pobre feliz porque se conforma con poco. Medir la felicidad o el bienestar a través de la riqueza no resulta acertado ya que nos interesa la posición económica en términos de mantener un flujo de consumo y disfrutar de cierto nivel de vida, no valorar el stock acumulado de activos. Riqueza y renta están obviamente relacionados, ya que mayor riqueza implica mayor capacidad de generar renta. Si escogiéramos consumo como indicador de bienestar, obtendríamos posiblemente resultados de distribución similares a los obtenidos con la renta, pero podemos encontrar casos en los que unidades consumen poco a pesar de contar con mucha renta y por tanto con mayores posibilidades (rico avaro o estoico). El consumo no nos permitiría en tal caso captar las verdaderas posibilidades del individuo o la familia.

El enfoque que se sigue en este capítulo 6 es introductorio, ello quiere decir que se incluye materia básica y se explican los métodos desde el nivel más elemental para que aquellos lectores que se enfrentan al estudio de la distribución y desigualdad de la renta por primera vez no tengan que contar con la preparación o lectura de material adicional previo. Además, se han tratado de explicar -mediante el texto y con apoyo de los ejercicios- cuestiones relevantes para el análisis empírico, pero accesibles a los lectores cuenten o no con formación económica. Tampoco es necesario tener formación matemática avanzada (pero sí la elemental en un nivel de secundaria) para comprender el texto o los ejercicios.

Para asimilar los conceptos que se introducen en los sucesivos epígrafes, el lector puede resolver todos los ejercicios que se proponen (y de los que se aporta solución) mediante la hoja de cálculo. Los datos que sirven para resolver todos los ejercicios corresponden a rentas en «unidades monetarias» cualquiera de 100 familias de distinto tamaño y composición. Los datos se han simulado para poder presentar todas las cuestiones de interés, pero los métodos de resolución empleados son igualmente válidos y aplicables a distribuciones de renta reales.

Tras la asimilación de los conceptos de este capítulo, el lector será capaz de transformar una distribución de renta corrigiendo las rentas por tamaño familiar según distintas escalas de equivalencia. Se incorporarán los concep-

[1] Las funciones de utilidad y e bienestar (agregaciones éstas de las anteriores) se utilizan básicamente para comparar situaciones entre individuos o colectividades, por lo que no importa tanto el valor alcanzado por las funciones (cardinalidad) como qué situaciones arrojan utilidad o bienestar mayor o menor, es decir, que lo realmente relevante es la ordinalidad de los resultados para su comparación.

tos de curva de Lorenz (como forma alternativa de la tradicional distribución de renta mediante histogramas) y se podrá ofrecer un valor de la desigualdad de la renta mediante estadísticos básicos o mediante índices de desigualdad específicos (se deriven o no de la curva de Lorenz). El lector también aprenderá a calcular índices de desigualdad que incorporan preocupaciones por la desigualdad variables como el índice de Theil o de Atkinson.

Todas estas cuestiones se organizan en los siguientes epígrafes. Tras esta introducción, el segundo epígrafe se dedica al estudio de la distribución de la renta y se repasa el contenido y cálculo de estadísticos básicos de las distribuciones, que son especialmente útiles en el caso de medición de la desigualdad de la renta. Se introduce el concepto de percentil como herramienta de partición básica de los colectivos con diferente renta y se termina el epígrafe con la explicación de las escalas de equivalencia, elemento clave en la corrección por tamaño y composición de la renta familiar. El tercer epígrafe se centra en el estudio de la curva de Lorenz, que constituye una herramienta básica en el análisis distributivo, y de la que se derivan algunos de los índices básicos de medición de la desigualdad. El cuarto epígrafe se centra en la explicación de algunos índices de desigualdad, primero los derivados de la curva de Lorenz: coeficiente de Gini y de Schutz, y después se hace referencia a la familia de entropía generalizada, y en particular al índice de Theil. Incluso este último índice se explica sin hacer referencia a las funciones de bienestar social, que se introducen en el quinto epígrafe para poder comprender el último de los índices que incorpora consideraciones de eficiencia y equidad, el índice de Atkinson. El sexto epígrafe ofrece algunas direcciones de sitios web de interés y concluye.

Los ejercicios propuestos se presentan en el propio texto, sin reservar un apartado final, con la idea de que se realicen en el momento adecuado: una vez que se han adquirido los conocimientos necesarios, y en el orden correcto: los ejercicios deben solucionarse en el orden en el que se proponen, ya que la resolución de los anteriores es necesaria para la comprensión de los siguientes.

6.2. DISTRIBUCIÓN DE LA RENTA

Cuando se cuenta con información de renta acerca de un colectivo de forma desagregada, ya sean individuos o familias, decimos que tratamos con micro-datos. La ventaja de contar con datos individualizados es que se puede realizar el análisis que se desee, clasificando y agrupando la información como más convenga al investigador. Al contrario, cuando los datos no se hallan de forma desagregada hay información que puede resultar interesante pero imposible de recuperar.

Veamos un ejemplo. Supongamos que en un colectivo de 100 familias se sabe que la renta media es de 160 unidades monetarias, que la composi-

ción de los hogares es de 2 miembros en el 50% de los casos, 3 miembros en el 40% y el 10% restante son individuos solos. Si quisiéramos conocer la media de renta por composición de hogar ¿cómo procederíamos? Es imposible obtener la media por subgrupos, porque los datos no están desagregados. Si contásemos con la renta de cada una de las 100 familias, se podría agrupar por composición y obtener las medias de renta por cada tipo de hogar. En cambio, si supiéramos que las rentas medias de los hogares son 100 para los hogares de 2 miembros, 200 para los de 3 y 300 para los de 1, siempre podríamos obtener la media global a partir de las proporciones sobre la población total que representan cada uno de los subgrupos. Así, la renta media sería 0,5+100+0,4*200+0,1*300=50+80+30=160.

Es decir, que *a partir de los datos desagregados siempre se pueden obtener las agregaciones deseadas agrupando según la información disponible, pero si los datos ya están agregados conforme a un criterio determinado, posiblemente no se pueda obtener la información clasificada de otra manera.*

Con ello, lo que queremos destacar es que a ningún investigador le interesa tener micro-datos *per se*, sino por la posibilidad de agrupar la información de la manera más conveniente. Cuando se cuenta con micro-datos se realizan agregaciones y se manejan estadísticos descriptivos de la distribución porque sería inmanejable contar con todos los datos para describir la realidad. Agregar micro-datos para describir una distribución es la única forma de proceder si se quiere describir lo que ocurre, y lejos de perder información al agregar, lo que se consigue es hacer manejable una base de datos. La clave está en poder contar con los datos micro originales para poder realizar las agregaciones de la forma que se desee.

Cuando se construye una distribución de renta se ordenan las rentas de todas las observaciones desde la más pequeña hasta la más grande. Podemos hablar de una distribución de renta datos discretos (no continuos), es decir que entre un nivel de renta y el siguiente no tiene porqué haber valores intermedios que sean ostentados por alguna observación. Por ejemplo, la renta más baja puede ser nula y la siguiente en orden de 100 unidades monetarias, sin que haya nadie que cuente con ningún valor entre 1 y 99. Incluso aunque así fuera, y todos los números enteros estuviesen «ocupados» por alguna observación, habría infinitos valores entre dos cifras enteras, por lo que no podemos hablar estrictamente de valores continuos. Lo que ocurre, es que los datos discretos se «linealizan» en ocasiones tratando de aproximar la distribución discreta a una función continua que cumpla con propiedades matemáticas convenientes [2].

[2] En la medida de lo posible, mantendremos en este capítulo el enfoque discreto por ser más sencillo para nuestros fines, pero también es posible describir los estadísticos de una distribución a partir de un enfoque continuo.

6.2.1. Descriptivos básicos: media, mediana, máximo mínimo y medidas de dispersión

Si las descripciones no se realizan mostrando todos los datos disponibles, ¿cuáles son los estadísticos básicos que describen una distribución de renta?. Podemos contar con varios, vamos a explicar el concepto de media, mediana, moda, máximo, mínimo, y algunas medidas de dispersión, haciendo hincapié en la interpretación de cada uno de ellos.

La **renta media** es la suma de todas las rentas dividida entre el número de observaciones. Si hay N familias o individuos con renta x_i, la renta media μ, se calcularía como sigue:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad i = 1,2,3,...N$$

La renta media se puede interpretar como la renta que tendría cada unidad si repartiésemos la renta total a partes iguales. Es posible que ninguna observación cuente con una renta igual a la renta media, pero este estadístico es informativo de acerca de si cada unidad se encuentra con la distribución existente mejor o peor que con una distribución uniforme en la que todas las rentas fuesen idénticas.

EJERCICIO 1: Obtenga la renta media de las 100 familias supuestas en el fichero de apoyo.

SOLUCIÓN: (Se aporta en la casilla B104). Puesto que las 100 rentas familiares se sitúan desde la casilla B2 hasta la B101, se realiza la suma de todas ellas SUMA(B2:B101) y se divide entre el número de familias, que son 100. El resultado es 7872,34 unidades monetarias. Si se comprueban los datos de las rentas familiares, no hay ninguna que cuente con esa renta exactamente, las familias 64 y 65 cuentan con renta inferior y superior a la media respectivamente. La información que nos aporta la media es que si repartiésemos la renta total de las 100 familias a partes iguales, hasta la familia número 64 saldrían ganando (ya que su renta actual es inferior a la media) mientras que solamente de la 65 en adelante saldrían perdiendo. Como nuestra muestra cuenta con exactamente 100 observaciones, ello implica que el 64% de las familias cuenta con una renta inferior a la media, y el 36% restante, con una renta superior. Aunque de forma muy precaria, pero nos proporciona una idea de la desigualdad en la distribución.

En realidad en nuestro ejemplo no hemos tenido que ponderar las observaciones por la frecuencia que representan en el total (f_i), ya que supo-

nemos que cada renta de nuestro ejemplo se presenta una de cada 100 veces. Si cada familia representase a una proporción del total pero no de forma uniforme (como ocurre en nuestro ejemplo 1/100), la media aritmética de la renta se obtendría por la suma de todas las rentas ponderadas por lo que cada familia representa respecto al total:

$$\text{Media} = \mu = f_1 x_1 + f_2 x_2 + \ldots f_N x_N = \sum_{i=1}^{N} f_i x_i$$

La **renta mediana** es aquel nivel de renta de la distribución que separa a las observaciones de la distribución en dos grupos numéricamente iguales, es decir el nivel de renta por encima del cual el 50% de la población es más rica y el 50% más pobre.

EJERCICIO 2: Calcular la renta mediana de la distribución y comprobar si es superior o inferior a la renta media.

SOLUCIÓN: (Se aporta en la casilla B103). La mediana separa la distribución en dos partes iguales. Como contamos con 50 observaciones, la mediana será la renta intermedia entre la observación 50 y 51 (marcadas en amarillo). Tales rentas son 5754 y 5798. La diferencia entre ambas es 5798-5754=44, luego la mitad serían 22 unidades monetarias, que sumadas a 5754 proporcionan una mediana de 5776. Excel calcula automáticamente la mediana de un grupo de observaciones sin más que indicar el rango en el que están situadas: MEDIANA(B2:B101).

En este caso, la renta media se sitúa por encima de la renta mediana, lo que revela una asimetría hacia la derecha (o positiva) de la distribución

También se suele utilizar como estadístico de tendencia central además de la media y la mediana, la moda, que es el valor más frecuente. En nuestro ejemplo todas las rentas se presentan con la misma frecuencia, por lo que no tiene sentido calcularlo.

El **máximo** y el **mínimo** en una distribución de renta indican cuál es la renta del individuo o la familia más rica y más pobre respectivamente. Cuando se cuenta con datos de una distribución de renta es interesante comprobar cuáles son estos valores para corroborar que son coherentes con lo esperado. Por ejemplo, la renta mínima en ocasiones aparece con valor negativo, y esto es perfectamente posible si se trata de renta neta, después de haber descontado impuestos o algún otro pago. Ahora bien, si se trata de renta salarial, no tendría sentido encontrar valores negativos. Con respecto al valor máximo de la renta, hay que decir que puede tratarse de un valor muy distante de las observaciones anteriores, ya que las grandes

diferencias en una distribución aparecen en la cola alta, es decir, en las rentas más elevadas.

EJERCICIO 3: Localizar el máximo y el mínimo de las 100 rentas familiares utilizando las funciones de Excel.

SOLUCIÓN: Hallar las rentas máximas y mínimas en este caso en el que solamente contamos con 100 observaciones y además ordenadas es obvio, ya que implica mirar la primera y última de las observaciones. Pero cuando se cuenta con ficheros de muchos datos que además no se encuentran ordenados, puede ser útil contar con las funciones de búsquedad de máximo (en la casilla B105: MAX(B2:B101)) y mínimo (en la casilla B106: MIN(B2:B101)).

Otros estadísticos descriptivos útiles para conocer la distribución de la renta son los que se refieren a la dispersión de los datos. Los datos más dispersos o más separados están indicando mayor desigualdad de la renta. Pero esta desigualdad debe ser independiente de las unidades monetarias en que se mide la renta, por ello hay medidas más y menos convenientes.

Una primera medida de dispersión es el **rango de variación**, que calcula la diferencia entre la renta máxima y la mínima. Pero este estadístico depende de las unidades de medida, es decir no es invariante ante cambios de escala. Veámoslo con un ejemplo: comparamos el rango de variación de las rentas de las familias si se utilizan las unidades monetarias actuales (supongamos euros) y si se utilizasen euros (lo que implicaría multiplicar por 167):

Rango de variación en euros = (50345-10)=50.335

Rango de variación en euros =50335*167=8.405.945

Como se puede ver, en función de las unidades monetarias en que se exprese la distribución y por tanto el rango de variación, la idea de dispersión puede ser una u otra. De igual modo, si expresásemos la renta en millones de euros, la dispersión sería de solamente 0,005.

Aportamos otras medidas frecuentemente utilizadas como la varianza, la desviación típica, la desviación media y el coeficiente de variación.

La **varianza** es otra medida de dispersión que se define como la media de todas las dispersiones al cuadrado de cada observación con respecto a la media de la distribución.

$$\text{Varianza} = \sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} \; i = 1, 2, \ldots N$$

Tampoco es invariante ante cambios de escala, ya que si las rentas se multiplican por k, la nueva varianza se multiplica por k^2, ya que se trata de una medida cuadrática. La **desviación típica** es la raíz cuadrada de la varianza, y tampoco es invariante ante cambios de escala, ya que si las rentas se multiplican por k, la desviación típica lo hace por ese mismo número.

$$\text{Desviación típica} = \sigma = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}}$$

La **desviación media** es la media de la suma de todas las diferencias en valor absoluto de cada observación con respecto a la media. Se toma el valor absoluto porque al haber rentas superiores e inferiores a la media, las diferencias toman valores positivos y negativos que podrían compensarse entre sí y no captar la dispersión.

$$\text{Desviación media} = DM = \sum_{i=1}^{N} \frac{|x_i - \mu|}{N} \; i = 1, 2, \ldots N$$

La desviación media se multiplica por k si todas la observaciones se multiplican por ese mismo número.

El coeficiente de variación es otra medida de dispersión que permite evitar los problemas de la invarianza ante cambios de escala, ya que no es sensible a las unidades de medida. Esta medida de dispersión es entonces muy útil cuando se quieren comparar distribuciones de renta que no están expresadas en las mismas unidades monetarias. Se define como el cociente entre la desviación típica y la media de la distribución.

$$CV = \frac{\sigma}{\mu}$$

EJERCICIO 4: Calcular sobre la distribución de renta de las familias las medidas de dispersión anteriormente descritas: rango, varianza, desviación típica, desviación media y coeficiente de variación. Calcularlas igualmente sobre la renta duplicada y comentar los resultados.

SOLUCIÓN: Para la distribución de rentas original los resultados se muestran en la columna B y para la duplicada, en la D. El rango de variación se obtiene por diferencia entre la renta máxima y mínima, se calcula en la casillas B107 y D107. La varianza se calcula con la función VAR(B2:B101) o VAR(D2:D101), y se presenta en las casillas B108 y D108. La desviación típica es la raíz cuadrada de la varianza, y se obtiene como =B108^0,5 o bien

D108^0,5 y se muestra en las casillas B109 y D109 . También existe una función pre-programada de Excel para obtenerla: DESVEST, como se explica en el comentario de la casilla A109. La desviación media se calcula en las casillas B110 y D110 y se utiliza DESVPROM(B2:B101) y DESVPROM(D2:D101). El coeficiente de variación se obtiene a partir del cociente de la desviación típica y la media (B109/B104) y (D109/D104) en las casillas B111 y D111. La tabla siguiente resume los resultados:

CUADRO 2.1.1. *Solución al ejercicio 4: Comparativa de estadísticos descriptivos de la distribución inicial y duplicada*

Estadísticos descriptivos	*Rentas originales*	*Rentas duplicadas*
Mediana	5.776	11.552
Media	7872,34	15744,68
Máximo	50.345	100.690
Mínimo	10	20
Rango	50.335	100.670
Varianza	74316386,28	297265545,1
Desviación típica	8620,69	17241,39
Desviación media	5260,98	10521,96
Coeficiente de variación	1,09506	1,09506
N	100	100

Al duplicar las rentas se duplican la mediana, media, máximo, mínimo, rango, desviación típica y desviación media. La varianza se multiplica por 4 (2^2) por tratarse de una medida cuadrática. Sin embargo, el coeficiente de variación es el único de los estadísticos descriptivos que permanece inalterado.

Ello no quiere decir que solamente el coeficiente de variación sea una medida válida de dispersión. La varianza y desviaciones son muy utilizadas porque informan acerca de los alejadas que se encuentran las observaciones del valor medio. Además, cuando no se varían las unidades de media, las comparaciones pueden resultar muy útiles a partir de la varianza. Pero solamente el coeficiente de variación aportaría coherencia en la comparación de distribuciones de renta medidas en unidades diferentes.

6.2.2. Percentiles

En ocasiones las distribuciones de renta aparecen agregadas en grupos llamados percentiles. Esta es una denominación genérica que adquiere una nomenclatura u otra en función del número de grupos en que se parta la distribución total. Los percentiles son particiones que contienen al mismo número de observaciones, por tanto, si partimos la población total en cuatro grupos, cada uno de ellos contendrá al 25% de la población, y la denominación al haber cuatro grupo es de «cuartiles» o «cuartilas». Si el reparto se distribuye en cinco grupos de población, cada uno contendrá el 20% de la población y la denominación esn este caso es de «quintiles» o «quintilas». Otras particiones habituales son las decilas, que contienen cada una al 10% de la población o las centilas, que contienen al 1%. Agrupar las distribuciones de esta forma es muy útil para forjarse una idea de lo que ocurre sin tener que visualizar todas las observaciones. Por ejemplo, a partir de los datos de nuestro fichero de renta, podemos agrupar la información de la renta en diez grupos que contengan el mismo número de observaciones, por lo que habremos creado decilas.

EJERCICIO 5: Calcular la renta media por decilas y obtener de nuevo los estadísticos descriptivos, comparando con la información desagregada.

SOLUCIÓN: Las decilas se crean en la columna G tras dividir el identificador de familia de la columna A entre 10 (en la columna F) y redondear sin decimales en la columna G mediante la función REDONDEAR.MAS(F2;0). Se obtiene la media sumando las 10 observaciones resultantes en cada decila y dividiendo entre 10 (es importante no dividir entre 100, pues el número de familias por decila es de 10 aunque el total de observaciones de la distribución desagregada sea 100). En las columnas J y K se agrupan la información de la decila y la renta media por decila respectivamente. Aplicando los cálculos descritos en los ejercicios anteriores, obtenemos los estadísticos descriptivos básicos, que se comparan en el siguiente cuadro.

Como se puede comprobar, se ha dividido entre 10 el número de observaciones ya que se partía de 100 datos originales que se han agrupado en 10 iguales en número (N). El único estadístico que no se altera es la media, ya que la media por decilas se construye agregando la información de las 100 familias originales, lo que implica que no hay pérdida de información. Lo que ocurre con cualquiera de las medidas de dispersión es que bajan de valor en cualquier caso, así como el valor del máximo [3]. Al comparar las dos distribuciones (desagregada y agregada por decilas) y ver los cambios en los estadísticos, queda patente que los cambios en la distribu-

[3] El valor del mínimo se eleva, lo que refleja también menor dispersión.

ción alteran enormemente la varianza, pero no tanto la desviación típica y media o el coeficiente de variación. Para que el orden de magnitud de las medidas no desvirtúe la envergadura del cambio, se presenta el cambio porcentual en la última columna del cuadro (obtenido en la columna M). Ello pone de manifiesto que el cambio porcentual en la desviación típica es idéntico al experimentado en el coeficiente de variación, a pesar de que el valor elevado de le desviación típica (o bajo del coeficiente de variación) podrían llevar a concluir que varía más la desviación que el coeficiente.

CUADRO 2.2.1. *Solución al ejercicio 5: comparación de estadísticos descriptivos entre la distribución original y agregada por decilas*

Estadísticos descriptivos	*Distribución original*	*Distribución por decilas*	*Cambio porcentual*
Mediana	5.776	5.815	0,67%
Media	7872,34	7872,34	0,00%
Máximo	50.345	29.275	–71,98%
Mínimo	10	619	98,39%
Rango	50.335	28.655	–75,66%
Varianza	74.316.386,28	67.747.844,24	-9,70%
Desviación típica	8.620,69	8.230,91	–4,74%
Desviación media	5.260,98	5.252,097	–0,17%
Coeficiente de variación	1,09506	1,04555	–4,74%
N	100	10	

6.2.3. Escalas de equivalencia

Un asunto de especial relevancia cuando se calcula la desigualdad de la renta es la consideración de la comparación entre unidades familiares diferentes. Contamos por tanto con dos fuentes de diferenciación:

a) Las familias tienen rentas distintas.

b) Las familias están compuestas de forma diferente.

Si comparamos dos familias A y B que cuentan con rentas Y_A= 150 unidades monetarias e Y_B=140 respectivamente, posiblemente consideremos mejor la situación de A, que disfruta de mayor renta. Pero esto puede ser erróneo si no se tiene en cuenta la composición del hogar. Si la familia A se compone de dos adultos y ocho menores, mientras que la B está formada por un individuo soltero, las 10 unidades monetarias adicionales de las que disfruta posiblemente no compensen para un adulto más y ocho menores. Una forma de hacer las comparaciones entre rentas más coherentes es corregir de alguna forma las rentas de las familias en función de la composición de los hogares. La primera solución —tan extrema como no corregir las diferencias no debidas a la renta— sería dividir las rentas entre el número de miembros de la familia. Pero esta corrección es «excesiva» ya que dos personas de un hogar no necesitan el doble de renta que viviendo por separado. Esto sucede porque hay una serie de «costes fijos» en el hogar que se comparten: por ejemplo, cuando la luz se enciende ilumina a todos los que comparten una estancia, los electrodomésticos no se multiplican por el número de miembros (hay un solo frigorífico o lavadora en cada hogar), y si bien familias más grandes precisan más espacio, duplicar los metros cuadrados de una vivienda no implica duplicar el coste de la vivienda, ya que en las viviendas más grandes el precio del metro cuadrado es más reducido que en las pequeñas.

Veamos un ejemplo de distribución de renta inicial de los hogares de composición diferente y cómo se transformaría si se considerase la distribución de renta *per capita*. En la subhoja «Escalas de equivalencia» de nuestra hoja de cálculo de ejercicios presentamos en la columna E el cociente entre A y C, es decir la renta dividida entre el número de miembros. Si creamos un nuevo identificador correlativo en la columna F y posteriormente ordenamos de menor a mayor tanto la renta *per capita* como el identificador (columnas H e I), podremos saber cómo han variado las posiciones de las familias por el hecho de aplicar la corrección por número de miembros. En la columna J se obtienen los saltos que experimentan las familias. Un valor positivo indica que se adelantan en la posición de su renta respecto al resto, un valor negativo implica empeoramiento en el orden que ocupan, y un cero, que siguen ocupando el mismo lugar que cuando no se corrige la renta por tamaño del hogar. La suma de todos los saltos debe ser cero, ya que siempre que una familia adelanta una posición, otra debe retrasarse.

Veamos cómo se interpretan los saltos en un caso concreto. Las familias que ocupaban el lugar 22º y 23º según el orden de la distribución inicial contaban con 2518 y 2633 unidades monetarias, pero al corregir por 2 y 3 miembros en cada una de ellas se convierten en 1259 (2518/2) y 878 (2633/3) unidades monetarias. Ello hace que la familia 22º pase a ocupar la posición 8ª, retrasándose (8-22=-14) 14 posiciones y que la familia 23º ocupe el lugar 42, adelantándose (42-23=19) 19 puestos. Debe tenerse en cuenta que el ade-

lanto y retraso de las posiciones no depende exclusivamente de cómo cambie la renta de cada familia, sino también de lo que ocurra con las demás, puesto que hablamos de posiciones relativas. Ello explica que rentas que en principio no cambian porque corresponden a una familia de un solo miembro (véase la observación 90) retrasen incluso 17 posiciones.

Hasta ahora hemos presentado las dos situaciones más extremas en cuanto a aplicación de escalas de equivalencia: no corregir y ponderar cada miembro del hogar al 100%. Pero como hemos anticipado, existen economías de escala por vivir en familia y compartir gastos, y para hacer comparables las rentas de los hogares de N miembros, no es necesario dividir entre N, sino entre algo más pequeño que N. Pensémoslo así: si pasar de ser un hogar de 1 a un hogar de 3 no triplica los gastos y las necesidades de renta, ya que se elevan los gastos en un factor inferior a 3, si comparásemos la renta de la que disponemos siendo 3 con la que disfrutábamos en solitario tendríamos que dividir entre algo menor que tres. Es decir, pasar a ser una familia de 3 no implica disponer *per capita* de la tercera parte de la renta sino de una proporción mayor.

Esta es la idea subyacente a las escalas de equivalencia, cuya aplicación es siempre controvertida por la subjetividad de que son acusadas. La clave está en por cuánto debemos ponderar a cada miembro del hogar, y esa ponderación se aplicará de forma uniforme a toda la distribución de renta. Pero en cada hogar los miembros suponen mayor o menor carga en términos económicos y ante la imposibilidad de tener en cuenta las peculiaridades de todos los hogares, lo que se suele hacer es discriminar exclusivamente entre adultos y menores.

Según la clasificación de Mancero (2001) las escalas de equivalencia pueden agruparse en cuatro categorías:

1) *Escalas de comportamiento:* se estiman a partir del gasto observado de los hogares. La idea inicial se basa en comparar el gasto necesario g, para a alcanzar determinado nivel de utilidad u, dados los precios en un hogar de composición z^h. ($g=g(u, p, z^h)$). Alcanzar esa misma utilidad a esos mismos precios costaría otra cantidad ($g=g(u, p, z^0)$) si el hogar estuviese compuesto de otra forma (z^0). Si tomamos como referencia el hogar de composición z^0, la escala de equivalencia se obtendría del cociente entre los dos gastos:

$$\text{Escala}= g(u, p, z^h)/ g(u, p, z^0)$$

Existen distintos métodos para estimar escalas de esta forma, pero las diferencias residen en cómo se estimen las demandas [4].

[4] Nótese como a pesar de que la función de utilidad no es conocida, a partir de las propiedades de la función indirecta de utilidad y la función de gasto, es posible estimar demandas.

2) *Escalas paramétricas:* se calculan a partir de una forma funcional, con parámetros explícitos que reflejan el grado de economías de escala y la «equivalencia por unidad de consumidor» de los miembros del hogar.

3) *Escalas expertas:* se construyen en base al criterio de investigadores utilizando información de distinto tipo y usualmente teniendo en consideración el uso específico que se le dará. Las escalas expertas-estadísticas pretenden únicamente contar personas, mientras que las expertas-programáticas se usan para asignar correctamente beneficios en los programas sociales.

4) *Escalas subjetivas:* se estiman a partir de la percepción subjetiva de las personas sobre sus necesidades y los gastos necesarios según composición demográfica. Se obtienen tras encuestar a las familias acerca de la cuantía de recursos que estiman necesaria para cubrir los gastos mínimos (lo que lleva aparejado un nivel de utilidad mínimo). El problema es que las respuestas están muy condicionadas por la renta familiar, de manera que familias más ricas declaran mínimos de renta más elevados para sobrevivir.

Para una primera aproximación en el uso de las escalas de equivalencia, nos centraremos en una de las escalas más utilizada dentro de las paramétricas, la de Oxford o de la OCDE. Las escalas paramétricas son muy utilizadas en el trabajo empírico, y si bien adolecen de falta de sustento teórico, son muy fáciles de aplicar. También explicaremos la escala de Buhmann *et al* (1988) que es una escala paramétrica y clasiflicable también como experta. Estas escalas (como cualquiera otra que se aplique) convertirán la renta inicial en una cantidad igual (si el hogar es unipersonal) o menor (si consta de más de un miembro). Al dividir la renta original por la escala, la renta está corregida, y se habla entonces de «renta equivalente». La ordenación de rentas según se corrijan o no por escalas de equivalencia será la siguiente:

Renta sin corregir >= Renta equivalente >= Renta per capita.

$$Y \quad >= \quad Y/E \quad >= \quad Y/N$$

Escala de la OCDE:

La escala de la OCDE pondera de forma diferente a adultos y niños en el hogar. Para dejar invariante la renta en el caso de los hogares formados por un solo adulto, el primer adulto se pondera por 1. Los adultos sucesivos se ponderan por 0,7, y los menores (considerados aquellos de menos de 14 años) se ponderan por 0,5. Este menor peso otorgado a los

menores asume implícitamente que el mantenimiento de un menor es menos costoso que el de un adulto. La escala se escribe considerando A el número de adultos y N el de menores, como:

$$\text{Escala OCDE} = 1 + 0{,}7(A-1) + 0{,}5N$$

En ocasiones se modifican los parámetros de esta escala, y la llamada escala modificada de la OCDE pondera por 0,5 a cada adulto además del primero y por 0,3 a los menores:

$$\text{Escala OCDE modificada} = 1 + 0{,}5(A-1) + 0{,}3N$$

Escala de Buhmann:

La escala de Buhmann *et al.* (1988) no distingue entre adultos y menores, toma el número total N de miembros del hogar y lo eleva al parámetro q, comprendido entre cero y uno, que se conoce como «elasticidad de equivalencia».

$$\text{Escala de Buhmann} = N^{q}$$

En el caso de que q tome valor nulo, es como si no se aplicase escala ninguna, ya que la renta no se modifica. Si q toma valor unitario, estamos ante la escala más «correctora» ya que esta situación equivale a calcular la renta *per capita*.

Una variante de esta escala que se ha utilizado en la construcción de líneas de pobreza en EEUU es la escala de dos parámetros Citro y Michael (1995), que pondera de forma diferente a adultos y menores.

$$\text{Escala Citro} = (A + pM)^{q}$$

Los adultos quedan ponderados por 1, mientras que los menores se ponderan mediante $p<1$, de forma menor. Los autores recomiendan escoger valores de p entre 0,65 y 0,75 y tomar q igual a 0,7 para así obtener resultados coherentes con otras escalas de equivalencia.

EJERCICIO 6: Corregir la renta de las familias según la escala de la OCDE, OCDE modificada, Buhmann (q=0,7), Citro (q=0,7 y p=0,65) y renta *per capita* y comentar los resultados de las rentas medias obtenidas y las desviaciones típicas de las distribuciones. Para ello suponga que en los hogares de 2 miembros ambos son adultos, y que por encima de dos, todos son menores. Compruebe que modificando la elasticidad de equivalencia en la escala de Buhmann en los casos extremos se obtiene o bien la misma renta, o la renta *per capita*.

SOLUCIÓN: El ejercicio 6 se resuelve en las columnas F a J entre las filas 104 y 208. Lo primero que se calcula es cuántos adultos y menores hay en cada hogar a partir de ala asunción de que no existen familias monoparentales. El número de adultos se obtiene en la columna C considerando que si hay más de dos miembros en el hogar, hay dos adultos, y si no, solo habrá adultos en el hogar(=SI(B107>=2;2;B107)). El número de menores se obtiene por diferencia entre el total de miembros y número de adultos.

En la columna F se calcula la renta equivalente aplicando la escala de la OCDE. La expresión de cálculo es (=E107/(1+$F$104*(C107-1)+$F$105*D107), donde E recoge las rentas sin corregir, que se dividen teniendo en cuenta nº de adultos (columna C) y menores (D) y la ponderación en las casillas F104 y F105 para adultos y menores respectivamente. Estas últimas casillas se bloquean con $ para poder copiar la fórmula a todas las casillas.

En la columna G, la expresión es la misma, excepto en las ponderaciones (=E107/(1+$G$104*(C107-1)+$G$105*D107)), que se sitúan en las casillas G104 y G105.

En la columna H se aplica la corrección por la escala de Buhmann, (=E107/(B107)^$H$104), siendo H104 la casilla que se inmoviliza por contener el parámetro q. Si se modifica de 0,7 a los valores extremos propuestos comprobamos que:

a) Si q (H104) es 0, la distribución coincide con la de la columna E, que contiene la renta inicial sin corregir

b) Si q (H104) es 1, la distribución coincide con la de la columna J, que contiene la renta *per capita*.

En la columna I se calcula la renta equivalente con la escala de Citro (=E107/(C107+$I$105*D107)^$I$104), en la que de nuevo se consideran adultos y menores por eparado, ponderándose los menores según el parámetro de la casilla I105, y donde I104 contiene la elasticidad de equivalencia.

De la observación de las medias y desviaciones típicas se concluye que efectivamente las distribuciones son muy diferentes según se apliquen o no correcciones por escalas de equivalencia, pero no varían tanto entre las escalas de equivalencia calculadas con los parámetros escogidos. Las rentas bajan en cualquier caso, porque en definitiva estamos «repartiendo» entre miembros del hogar, y asimismo lo hacen las desviaciones típicas, ya que cualquiera de las distribuciones de renta equivalente es menos dispersa que la inicial.

Entre las dos escalas de la OCDE, la no modificada da lugar a rentas menores, ya que asume menos economías de escala por vivir en el hogar que la modificada. Para los parámetros escogidos, las rentas equivalentes corregidas por Buhmann y la OCDE modificada son bastante similares (mucho más lo son las correcciones por Buhmannn y Citro, que difieren exclusivamente en la ponderación de los menores). También puede comprobarse la sensibilidad de las distribuciones ante cambios en los parámetros sin más que modificar los valores supuestos, por ejemplo fuera de los rangos recomendados, en cuyo caso las diferencias entre distribuciones son mucho más notables.

La mera visión de los datos solamente puede prevenirnos de que existen diferencias entre las rentas originales y las corregidas por escalas de equivalencia, pero donde realmente se hará patente el efecto de las mismas en cuando llevemos a cabo análisis de la desigualdad de la renta, cuestión que abordamos inmediatamente en el apartado 3.

6.3. LA CURVA DE LORENZ

Una herramienta tremendamente útil para representar la distribución de la renta (o de la renta neta, o cuotas impositivas, o gasto, o cualquier variable que se desee)es la curva de Lorenz. Esta representación de una distribución se realiza en un cuadrado de lado uno, donde en el eje horizontal o de abscisas se representa el tanto por uno de población acumulada, y en el eje vertical, el tanto por uno de renta acumulada. Para graficar la curva de Lorenz de la renta, es preciso haber ordenado previamente las observaciones de menor a mayor renta, de manera que la primera será la familia más pobre, y en último lugar, se hallará la más rica. Pongamos un ejemplo: supongamos cuatro familias con rentas 10, 15 60 y 100 respectivamente. Si tratásemos de construir la curva de Lorenz habría que calcular la proporción de renta que se va agregando al sumar familias sucesivas:

CUADRO 3.1. *Valores para cálculo de curva de Lorenz a partir de cuartiles*

	Acumulado de población (en tanto por uno)	*% Acumulado de renta (en tanto por uno)*
Familias=1	1/4=0,25	10/185=0,05
Familias=1+2	2/4=0,5	10+15/185=0,14
Familias=1+2+3	3/4=0,75	10+15+60/185=0,46
Familias 1+2+3+4 (todas)	4/4=1	10+15+60+100/185=1

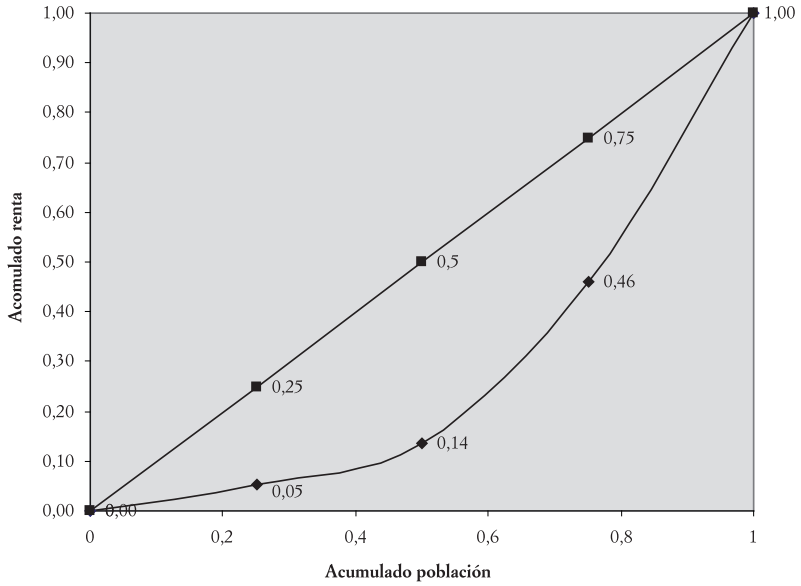


GRÁFICO 3.1. *Curva de Lorenz a partir de cuartiles*

Como se puede apreciar en el gráfico, la curva de Lorenz compara la proporción sobre el total de renta que acumulan las distintas proporciones de familias. Computar exclusivamente a la primera familia, la más pobre, implica incluir al 25% de la población (0,25 en tanto por uno) pero esa familia disfruta solamente del 5% de la renta total de las cuatro familias consideradas, que suma 185 unidades monetarias. Si pasamos a incluir una familia más, habremos computado al 50% de la población (0,5 en tanto por uno) pero esas dos familias no disfrutan del 50% de la renta, sino de menos, 14%. Al incluir a la tercera familia contamos con 0,75 de la población que cuenta con 0,46 de la renta. Por último, computar a las cuatro familias implica incluir la proporción total (1) de población y sumar asimismo toda la renta.

Destaquemos algunas características de la curva de Lorenz:

1) La curva de Lorenz comienza siempre en (0,0) porque cuando no se incluye ninguna familia tampoco se acumula ninguna renta, y termina en (1,1), porque toda la población posee toda la renta.

2) La curva de Lorenz permite de un vistazo comprobar si la renta se distribuye o no de forma desigual. Para ello se toma como referencia la diagonal o bisectriz, que marca la situación en que cada percentil de población disfruta exactamente de ese percentil de renta. Cualquier curva que se sitúe por debajo implica que no todas las unidades cuentan con la misma renta, y que el percentil x % de población disfruta de menos del x % de la renta.

3) La situación de igualdad máxima (todos las familias cuentan con la misma renta) implica una curva de Lorenz coincidente con la bisectriz.

4) La situación de desigualdad máxima (una familia cuenta con toda la renta y las restantes no tienen nada) originaría una curva de Lorenz coincidente con los ejes, sería cero hasta agregar a la última familia y en ese punto se elevaría hasta 1.

5) La curva de Lorenz nunca puede situarse por encima de la bisectriz, ya que al haber ordenado por renta y graficar renta acumulada, como máximo se puede dar una situación de igualdad perfecta y que el x % de la población disfrute del x % de la renta, pero no de más!?

6) Cuando se representan dos distribuciones de renta alternativas con la misma media sobre la misma población y no se producen cruces entre las curvas de Lorenz, se puede asegurar que la curva que se sitúa por encima (se dice que «domina en sentido de Lorenz») lleva aparejada un mayor bienestar. Obviamente, situarse por encima quiere decir que la desigualdad es menor.

La expresión matemática de la curva de Lorenz en términos continuos implica el uso de integrales, lo que facilita las cosas cuando se trata de llevar a cabo análisis teórico, pero las complica en el campo empírico, ya que requeriría la estimación de la forma funcional de la distribución de la renta. Como nuestro interés reside en el aprendizaje del manejo de esta herramienta, utilizaremos la expresión en términos discretos, que implica el uso de sumatorios y supone la generalización del ejemplo explicativo.

Si se considera que las unidades cuya renta se acumula son $i=1,2,....N$, y que su renta es $x_i=x_1, x_2, ...x_N$, entonces el percentil acumulado hasta j , p_j sería:

$$p_j = \sum_{i=1}^{j}$$

Y la renta acumulada que le correspondería sería:

$$x_j = \sum_{i=1}^{j} x_i$$

[5] Si las rentas no estuviesen ordenadas, al tomar la familia más rica ocurre que representa al 25% de la población y disfruta de 100/185 (es decir, 54%) de la renta. Por eso las curvas de concentración –que representan una variable acumulada por un criterio distinto de ordenación– sí que pueden situarse por encima de la bisectriz. Véase más adelante citar capítulo de Jorge Onrubia.

Mientras que el total de población vendría dado por p_N y la renta total por x_N.

La curva de Lorenz para la población acumulada hasta el nivel *j* se escribiría como:

$$L\left(\frac{p_j}{p_N}\right)=\frac{\sum_{i=1}^{j} x_i}{x_N}$$

Con esta información debemos ser capaces de construir cualquier curva de Lorenz dada una distribución de datos desagregados. En el siguiente ejercicio se propone el cálculo de curvas de Lorenz en diferentes escenarios.

EJERCICIO 7: Construir la curva de Lorenz de la distribución inicial de la renta utilizando los datos referentes a las 100 familias.

SOLUCIÓN: En la sub-hoja «Curvas de Lorenz» se presentan los resultados en las columnas D y E entre las filas 1 y 102. Lo primero que se necesita es **ORDENAR** la renta. En nuestro caso la renta ya estaba ordenada, pero es preciso no olvidar hacerlo, ya que de otro modo no se construye una curva de Lorenz sino de concentración. Un avez ordenado se calculan los tantos por uno acumulados de población y de renta.

a) Cálculo de los tantos por uno acumulados de población: esta parte es más sencilla ya que al contar con u identificador de familias correlativo del tipo 1, 2, 3,...100, este valor nos indica cuántas familias se han ido acumulando, por lo que no hay más que dividir el identificador entre el total de familias. Ello se calcula en la columna D mediante =A2/$A$101, ya que en la casilla A101 se incluye el total de familas que se fija para copiar la fórmula desde D2 hasta D101.

b) Cálculo de los tantos por uno acumulados de renta: se obtiene en la columna E; para la primera familia, se divide la renta de esa familia entre el total de renta de las 100, que se calcula en la casilla B102. Par el resto de familias se suma la proporción acumulada hasta la casilla anterior y la proporción que añade esta nueva familia mediante la expresión: =E2+B3/$B$102, que se copia hasta el final.

c) Obtención del gráfico: se marca el rango de datos y se escoge la opción «dispersión». El resto de opciones y retoques se dejan a gusto del lector.

Una cuestión interesante sería llevar a cabo una comparación entre las distribuciones de renta corregidas por distintas escalas de equivalencia. Para ello escogemos las distribuciones calculadas en el apartado de escalas de equivalencia y calculamos sus curvas de Lorenz como se propone en el ejercicio 8.

EJERCICIO 8: Elaborar en un mismo gráfico las curvas de Lorenz de la renta inicial, de la renta corregida por las escalas de equivalencia de la OCDE, OCDE modificada, Buhmann, Citro y la renta *per capita*. Comente los resultados obtenidos.

SOLUCIÓN: La solución a este ejercicio se calcula en la sub-hoja de curvas de Lorenz entre las filas 105 y 206. El procedimiento es idéntico al descrito en el ejercicio 7, si bien es necesario **ORDENAR** de menor a mayor las rentas equivalentes, ya que al haber aplicado las escalas se producen saltos en todo caso. Los datos se pegan como valor (sin arrastrar la fórmula) entre las columnas A y G. Hay que obtener el total de nuevo de cada una de las rentas equivalentes (fila 206) y después se repite el proceso del ejercicio 7 entre las columnas J y N.

Una vez obtenido el gráfico, aunque existen tramos coincidentes, queda patente que la curva de Lorenz sin corregir por escala de equivalencia es la más cercana a la bisectriz, y la obtenida a partir de la renta *per capita*, la más alejada o la que muestra más desigualdad. Recuérdese que la distribución de partida es la misma, pero se ha corregido por composición y/o tamaño familiar. El caso de escala más extrema es la corrección dividiendo entre el número de miembros, por ello entre los dos extremos quedan comprendidas todas las curvas de Lorenz de las distribuciones de diferentes rentas equivalentes. Es lógico que al dividir la renta entre el número de individuos, las familias de mayor tamaño estén peor consideradas, lo que en términos de curva de Lorenz significa situarse más adelante en el cálculo de las proporciones de población y agregar rentas más pequeñas, lo que a la fuerza implica valores de la curva más bajos omás lejanos a la diagonal para cada percentil considerado.

Como se ve, ésta es una forma muy sencilla de mostrar las distintas distribuciones de renta corregida por escalas de equivalencia, y que permite comprobar la desigualdad medida en cada una de ellas.

6.4. ÍNDICES DE DESIGUALDAD

Como se ha mostrado en el epígrafe anterior, el uso de curvas de Lorenz constituye una forma muy sencilla de mostrar las distintas distribuciones de renta, y que permite comprobar la desigualdad derivada de cada

una de ellas. Pero esta simplicidad nos puede confundir, porque es necesario cuantificar el valor de la desigualdad a partir de índices, es decir, recoger en una sola cifra numérica toda la información referida da desigualdad que se puede extraer de la distribución. En muchas ocasiones se producen cruces en las curvas de Lorenz, pero aún así es posible calcular índices de desigualdad que nos informen acerca de cuál de las distribuciones comparada es más o menos igualitaria, o simplemente cuantificar en términos numéricos la desigualdad.

Dado que la curva de Lorenz es un instrumento que se utiliza con mucha generalidad en el análisis distributivo, también son de uso muy común los índices que a de ella se derivan, como el índice de Gini y el coeficiente de Schutz (este último se utiliza mucho menos que el Gini). Pero existen otros índices como la familia propuesta por Theil de índices de entropía generalizada que provienen de la Física y tratan de medir la desigualdad como el «desorden» existente en una distribución. El presente epígrafe explica estos índices que miden exclusivamente desigualdad, y en un apartado posterior se presentará el índice de Atkinson que combina consideraciones de equidad y de eficiencia.

6.4.1. Índices derivados de la curva de Lorenz (Schutz, Gini)

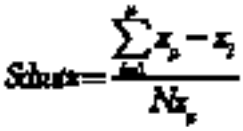
A partir de la construcción de la curva de Lorenz, se puede cuantificar la desigualdad en unción de lo separada que se encuentre la curva de Lorenz de la bisectriz, que marca la referencia de igualdad perfecta. Se puede medir la diferencia obteniendo la magnitud de todo el área entre la curva de Lorenz y la bisectriz, o bien medir la separación máxima entre bisectriz y la curva de Lorenz. La primera de las ideas es la base de la construcción del coeficiente de Gini [6], y la segunda es incorporada en el coeficiente de Schutz. Veamos cómo se calculan y la interpretación de cada uno de ellos.

Coeficiente de Schutz: Este coeficiente se obtiene como la máxima distancia de separación ente la curva de Lorenz y la bisectriz. La separación es máxima cuando las tangente a la curva de Lorenz es paralela a la bisectriz, es decir cuando la pendiente de la curva de Lorenz es de 45º y ello ocurre para el percentil que agrega observaciones hasta la media. Por tanto, si llamamos x_μ al nivel de renta medio y m al lugar ocupado en la ordenación por la observación que tiene la renta media, el coeficiente de Shutz se obtendría según nuestra notación como:

$$\text{Schutz} = \frac{p_\mu}{p_N} - \frac{\sum_{i=1}^{m} x_i}{x_N}$$

[6] Se denomina coeficiente de Gini al valor expresado entre 0 y 1, e índice de Gini si se multiplica por 100 este valor y se expresa en porcentaje.

Otra forma de obtener este coeficiente, que aunque es más engorrosa para los datos de los que disponemos, ofrece una intuición de su significado, es la siguiente:



En la expresión anterior, el numerador recoge todos los déficits de renta hasta la renta media de las observaciones que cuentan con renta inferior a la media. El denominador expresa la renta total. Su cociente está indicando el déficit de renta sobre el total de aquellos con renta inferior a la media. Por lo tanto, el coeficiente de Schutz indica la proporción de renta que habría que transferir desde las familias que tienen más de la renta media hacia las que están por debajo para que todas tuvieran lo mismo. Aunque la intuición que ofrece es buena para captar el grado de desigualdad, el coeficiente de Schutz adolece de un problema importante que hace que no se utilice con tanta generalidad como el coeficiente de Gini: cuando se producen transferencias entre rentas por encima de la media, o rentas por debajo de la media pero que no cruzan la media, el valor del coeficiente no varía. ¿Por qué esto es un inconveniente? En la medición de la desigualdad, habitualmente se comparan situaciones antes y después de determinada intervención pública como establecimiento de impuestos o pago de transferencias. Si se producen cambios en la distribución de modo que una familia más rica que otra paga impuestos que le son transferidos a la más pobre, sería deseable que el índice recogiese este cambio como una mejora en la igualdad, y por tanto debería descender de valor. Pero debido a que lo que una familia da es exactamente lo que otra recibe y no se ha alterado la media, el coeficiente de Schutz permanece inalterado. Esta misma sería la conclusión si las transferencia y el pago se hubiesen producido por encima de la media quedando ésta inalterada. Se dice por tanto que el coeficiente de Schutz no cumple el principio de las transferencias [7], según el cual un índice de desigualdad debe descender de valor cuando se transfiere renta de una unidad con más renta a otra con menos.

EJERCICIO 9: Calcular el coeficiente de Schutz para la distribución inicial y para la de rentas corregidas por escalas de equivalencia utilizando los dos métodos propuestos. Comente los resultados obtenidos.

a) Comenzamos por ser más sencillo el procedimiento, calculando la distancia máxima entre la curva de Lorenz y la bisectriz, referen-

[7] Cumpliría este principio solamente para transferencias que crucen (y por tanto alteren) la media, pero nuestro interés reside en que la propiedad se cumpla en todo caso, independientemente de entre quiénes se produzca, siempre que se transfiera de más ricos a más pobres.

te de máxima igualdad. En la hoja de ejercicios, se presentan los resultados en la sub-hoja llamada Schutz. Se toman los valores de las curvas de Lorenz y de la proporción que representan las familias y se paegan como valor en dicha hoja (columnas A a G). A continuación se calculan todas las diferencias de valor entre la diagonal y cada una de las curvas de Lorenz (columnas J a O). En la columna I volvemos a añadir el identificador de la familia para poder encontrar fácilmente la familia en la que se produce la máxima diferencia. Una vez calculadas las diferencias, en la fila 104 se obtienen todos los coeficientes de Schutz. Además se sombrea en azul el lugar ocupado por la familia donde la distancia es máxima, para comprobar que ocurre lo mismo cuando se utiliza el método alternativo.

b) Según la otra forma de cálculo propuesta, sería necesario calcular las rentas que se acumulan hasta la observación que contiene la renta media (en nuestro caso, al no tratar de de datos continuos, no hay ninguna familia cuya renta coincida con la media, pero agregamos hasta la que cuenta con la renta inferior, de modo que la familia con renta siguiente tendría más que la media y no se computaría). Se puede comprobar que mediante este sistema, las observaciones a computar coinciden con el método descrito en a). Los cálculos se realizan en las columnas J a O (para cada una de las distribuciones). En la fila 112 se identifica cuántas de las familias deben ser incluidas; en la fila 113 se suma la renta de las observaciones hasta el número especificado en 112; en la 114 se multiplica la media por ese número de veces. En la fila 115 se obtiene la diferencia entre las filas 114-113, lo que otorga el valor del déficit total por debajo de la media. En la fila 116 se recuerda el valor de la renta total en cada caso, dato con el que ya se contaba, y por último en la fila 117 se calcula el cociente del déficit por debajo de la media entre la renta total. Como se puede comprobar, los valores de los coeficientes son coincidentes en todos los casos.

De los resultados obtenidos se deriva coherencia con las representaciones de las curvas de Lorenz: si la curva más igualitaria correspondía a la de la renta sin corregir, también a esta distribución se le debe asociar el valor de la desigualdad menor, como efectivamente ocurre. La distribución más desigualitaria, con la curva de Lorenz más cercana a los ejes correspondía a la de la renta per cápita, cuyo valor de coeficiente de Schutz es más elevado. En medio de estos dos extremos se sitúan los coeficientes de las demás rentas corregidas por escalas de equivalencia. En orden de desigualdad mayor a menor, los resultados son:

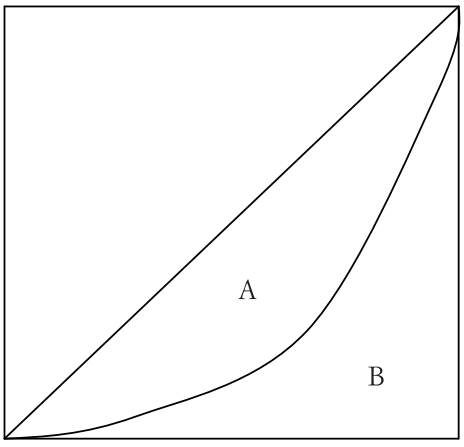
CUADRO 4.1.1. *Valor del coeficiente de Schutz en las diferentes distribuciones*

Renta per capita	*Escala OCDE*	*Escala Buhmann*	*Escala Citro*	*Escala OCDE modificada*	*Renta sin corregir*
0,4297	0,3909	0,3889	0,3761	0,3676	0,3341

Esta diferencia en la desigualdad, al partir de la misma distribución, se debe exclusivamente al hecho de que se consideren más o menos economías de escala por vivir en familia. Con respecto a la situación en que no se asumen economías de escala (renta *per capita*) las escalas que asumen más economías son en este orden: la de la OCDE, Buhmann, Citro, OCDE y modificada. (Obviamente para los parámetros de nuestro ejercicio, si éstos se modifican, la ordenación también cambiaría)

Otro de los índices de medición de la desigualdad que se deriva de la curva de Lorenz es el *índice de Gini,* en este caso, basado en medición de áreas y no en distancia como el coeficiente de Schutz. El índice de Gini se calcula a partir de la separación entre la curva de Lorenz y la bisectriz —referencia de igualdad perfecta— de manera que cuanto mayor es la separación entre ambas curvas, mayor es la desigualdad existente. Los valores máximos y mínimo del índice son 1 para el caso de desigualdad máxima y 0 cuando todas las unidades cuentan con idéntica renta. ¿Cómo se calcula en índice? Se calcula la proporción del área A en relación a A+B, es decir, el área de separación existente entre la curva de Lorenz y la referencia de igualdad perfecta (A) con respecto a la máxima separación que puede darse (A+B):

$$Gini=A/(A+B)$$



En el caso extremo en que una unidad poseyera toda la renta, la curva de Lorenz coincidiría con los ejes, presentando forma de L invertida, por lo que el área A coincidiría con A+B, y el vañor del índice de Gini sería (A+B)/(A+B), es decir, la unidad, presentando este valor la desigualdad máxima. Si al contrario, la renta estuviese igualitariamente distribuida, el área A sería nula, puesto que la curva de Lorenz coincidiría con la diagonal, presentando el índice de Gini un valor nulo (0/(A+B)), e indicando que no existe desigualdad. En cualquier otro caso en que la curva de Lorenz presente la forma habitual, el índice se sitúa entre 0 y 1 (normalmente toma valores en el entorno de 0,3 ó 0,4).

Debe notarse que dado que la curva de Lorenz se construye sobre ejes que varían de 0 a 1, el área del cuadrado en que se representa vale 1, y el triángulo A+B vale la mitad (0,5), por lo que también se puede expresar el índice como:

$$Gini = 2 \cdot A$$

ya que A/(A+B)=A/0,5=2·A. El índice es entonces el doble del área comprendida entre la curva de Lorenz y la bisectriz. La pregunta que surge automáticamente es por qué se define como el doble de esa área A, y no simplemente el área una sola vez. La respuesta es también sencilla: muchos de los índices de desigualdad se acotan entre 0 y 1 para hacer más fácil la comparación entre ellos, y de este modo, el Gini vale también 1 como máximo (y no 0,5 como sería el caso si no se duplicase el área A).

Existen numerosas expresiones alternativas para calcular el índice de Gini. Si se dispone de la forma funcional de la curva de Lorenz, no hay más que integrar el área que queda por debajo y se obtiene B. Pero habitualmente, el investigador se encuentra con datos «discretos» cuando se enfrenta a distribuciones de renta, por lo que es más operativo utilizar la expresión siguiente, que si bien parece compleja, resulta muy sencilla de aplicar en la práctica.

$$Gini = 1 + \frac{1}{N} - 2 \cdot \left(\frac{x_N + 2x_{N-1} + 3x_{N-2} + 4x_{N-3} + \ldots + (N-2)x_3 + (N-1)x_2 + Nx_1}{N^2 \mu} \right)$$

En la expresión anterior, *N* se refiere al número de unidades incluidas en la distribución, m se refiere a la renta media y las rentas x_N, x_{N-1}, ...x_1, son las rentas ordenadas de mayor a menor, de manera que la renta más elevada, x_N se pondera por el menor valor, 1, y la renta más baja, x_1, por el mayor valor, N, que es el número de observaciones.

Esta fórmula se obtiene de la medición del área A y B a partir de aproximaciones de triángulos y rectángulos cuando se cuenta con datos discretos. Recuérdese que al no contar con la forma funcional de la curva de Lo-

renz, lo que se obtiene es la aproximación poligonal a una curva, tanto más cercana a la curva real cuanto mayor sea el número de observaciones incluidas. Cuando se incluye toda la información acerca de renta de las unidades la aproximación al verdadero valor de la desigualdad es inmejorable, pero en muchas ocasiones, se calculan curvas de Lorenz a partir de datos agrupados en decilas, en los que existe un sesgo como se comprobará en los próximos ejercicios.

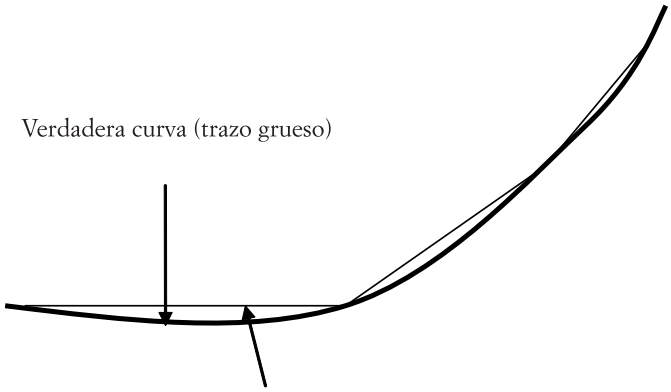
EJERCICIO 10: Obtener el índice de Gini a partir de la distribución de rentas familiar inicial utilizada en los ejercicios anteriores (sin corregir por escalas de equivalencia), y comparar el valor obtenido si se utilizasen exclusivamente los valores de las rentas medias de cada decila.

SOLUCIÓN: En la subhoja Gini se presentan los resultados a este ejercicio. En las columnas A y B se pegan los valores del identificador de familia y de la renta iniciales. El siguiente paso es construir un identificador inverso para ponderar de forma decreciente las rentas cuando están ordenadas de mayor a menor, de forma que en el caso en que se utilicen las 100 familias, la renta más baja se pondere por 100, la siguiente por 99, la siguiente por 98 y así sucesivamente, hasta la más alta ponderada siempre por 1. Para generar este identificador inverso no hay más que tomar el número de observaciones totales y restar el lugar que ocupa cada observación y añadir 1 (valor de la columna C=100-Ai+1). Después se multiplica la ponderación por las rentas en la columna D y se suman todos los valores en la casilla D102. La obtención del Gini es automática en la casilla D104, pues todos los valores necesarios se han obtenido ya: =1+(1/A101)-2*(D102/(D103*A101^2)). El Gini toma un valor de 0,4791.

Cuando se realiza el cálculo utilizando los valores medios de las decilas, es necesario tomar ciertas precauciones: la media será la misma porque al calcularse sobre percentiles, se cuenta con el mismo número de observaciones y las ponderaciones son coincidentes. Si se contase con 10 datos correspondientes a colectivos con distinto número de observaciones, aunque se tomase como N el valor 10, los resultados no serían coincidentes porque cada dato no representaría al 10% de la población (podría ser más o menos).

En las columnas G y H se toman los valores ya obtenidos en la subhoja de Estadísticos referentes a los valores medios de decilas. En la columna I se calcula el identificador inverso que servirá para las ponderaciones, teniendo en cuenta que solamente disponemos de 10 datos (y no 100 como antes). Las rentas medias por decilas se ponderan en la columna J, se suman en J12, y se cuenta con todos los datos para obtener el índice, en J14. En este caso el valor es de 0,4682, inferior al obtenido con los 100 datos.

Cuantos menos datos se utilizan, menor es el valor del índice obtenido, ya que la aproximación poligonal se produce por encima de la curva, y el área entre la curva de Lorenz aproximada y la diagonal es menor a la «real» o la que se obtendría con todos los datos. La siguiente figura puede ilustrar la idea que apuntamos:



En el ejemplo que hemos mostrado, la diferencia entre los índices de Gini obtenidos no es excesiva, si bien es cierto que cuando se utilizan datos reales, las diferencias pueden ser más grandes, sobre todo cuando se cuenta con una gran dispersión en la última decila (que es la que mayor desigualdad suele captar).

EJERCICIO 11: Obtener los índices de Gini para las rentas familiares corregidas con escalas de equivalencia de la OCDE y Buhmann y compararlas con el valor de la desigualdad obtenido para las rentas sin corregir.

SOLUCIÓN: En la subhoja Gini se obtienen los valores de los índices de desigualdad para las rentas equivalentes corregidas según la escala de la OCDE y Buhmann. Lo primero que se necesita es contar con las rentas equivalentes, que se obtuvieron en ejercicios anteriores. Es importante tener la precaución al pegar los datos de ORDENAR las rentas, ya que al aplicar las escalas de equivalencia, la ordenación no tiene porqué coincidir con la inicial (como es el caso). Una vez ordenadas, se aplica el mismo método descrito en el ejercicio 10, calculando de nuevo las medias, ya que también se modifican para cada renta equivalente. Los resultados se muestran en la casilla O104 para la renta corregida por la escala de la OCDE y T104 para la renta corregida por la escala de Buhmann con valores de 0,5398 y 0,5381 respectivamente. Los valores del índice son mayores en ambos casos con respecto al calculado sobre la renta sin corregir, lo que quiere decir que al considerar el «reparto» de las rentas familiares entre los miembros de la familia y las correspondientes economías de escala, la

desigualdad es mayor que sin aplicar tal corrección. Esto no tiene por qué ocurrir siempre, ya que depende de cómo se combinen la distribución de renta y de cargas familiares. Si las mayores cargas familiares se concentran entre la población de más renta, el efecto de aplicar escalas de equivalencia será el de disminuir la desigualdad, ya que las rentas más elevadas se «dividen más» mientras que las más bajas se quedan «más parecidas» a las rentas sin corregir, lo que a la fuerza disminuye la dispersión de la renta y por tanto la desigualdad. Si la concentración de cargas familiares mayores se produce en las rentas más bajas, el efecto será el contrario, aumentando la desigualdad.

6.4.2. Entropía generalizada: índice de Theil

Otra conjunto de índices diferentes, que no se derivan en este caso de la curva de Lorenz es la constituida por la familia de entropía generalizada, también conocida en el ámbito de estudio de la desigualdad como índice de Theil en un caso particular de la familia de entropía.

La expresión de cálculo del índice [8] es:

$$T=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{\mu}\cdot\ln\frac{x_i}{\mu}\right)$$

Donde x_i representa la renta del i-ésimo individuo o familia de las N totales, y μ es la renta media.

La interpretación es más sencilla si pensamos de nuevo en situaciones extremas. Si la renta se repartiese de forma uniforme, todo el mundo tendría una renta igual a la media, por tanto el índice de Theil valdría 0, ya que

$$T=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mu}{\mu}\cdot\ln\frac{\mu}{\mu}\right)=\frac{1}{N}\sum_{i=1}^{N}(\ln 1)=\frac{1}{N}0=0.$$

En el otro extremo, si una sola unidad contase con toda la renta y las demás no tuvieran nada, la renta del más rico sería $N\mu$ y el índice de Theil valdría $\ln(N)$, ya que en este caso $T=\frac{1}{N}\left(\frac{N\mu}{\mu}\cdot\ln\frac{N\mu}{\mu}\right)=\ln(N)$. La interpretación del índice de Theil no es tan intuitiva como la que obtuvimos para el Gini, pero se puede tratar de interpretar el sentido del mismo. El primer factor de la expresión representa el peso de una observación en la muestra

[8] Para un valor del parámetro a igual a uno, como se verá más tarde.

total, mientras que el segundo, el la proporción de su renta con respecto a la media. A pesar de su mayor dificultad de interpretación, el índice de Theil tiene la ventaja de ser descomponible, de manera que la desigualdad total puede ser obtenida a partir de la suma ponderada de las desigualdades de cada subgrupo. Por ejemplo, podría desagregarse la población de un país por regiones y obtener la desigualdad total del país a partir de la desigualdad existente en cada región ponderando por el peso de cada región en el conjunto del país. Si la población se subdivide en M grupos y s_k es la proporción de la renta total de la que disfruta de cada grupo k, si llamamos T_k al índice de desigualdad de Theil para la región k, y $\hat{I}k$ es la renta media del grupo k, tenemos que el índice de Theil se puede reescribir como:

$$T=\sum_{k=1}^{M} s_k T_k + \sum_{k=1}^{M} s_k \ln \frac{\mu_k}{\mu}$$

En realidad, hasta ahora hemos simplificado las cosas, pues el índice de Theil que hemos presentado es un caso particular de la familia antes mencionada de índices de entropía generalizada. El caso particular que se ha presentado implica que el valor del parámetro a es igual a la unidad. El caso general de índices de entropía generalizada (GE) se expresa como:

$$GE(\alpha)=\frac{1}{\alpha^2-\alpha}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{\mu}\right)^{\alpha}-1\right]$$

Si el parámetro a toma valor nulo, GE(0) será:

$$GE(0)=\frac{1}{N}\sum_{i=1}^{N}\frac{\mu}{x_i}$$

Si toma el valor 1, tenemos el índice de Theil (GE(1)=T), y otro valor habitual es a igual a 2 en cuyo caso el valor del índice GE sería:

$$GE(2)=\frac{1}{2N\mu}\sum_{i=1}^{N}(x_i-\mu)^2$$

El parámetro a representa la aversión a la desigualdad, de manera que cuanto mayor es el parámetro, menor es la preocupación por la desigualdad. Para entender este concepto, hay que hacer referencia al principio de las transferencias. Si existe preocupación en la sociedad por la desigualdad, se calificarán como mejores aquellas situaciones en las que se produce un reparto de rentas de los más ricos a los más pobres, sin cambiar la media de la distribución ni la renta total. Si esto es así, quiere decir que se consideraría favorable un reparto, o dicho de otra forma «se cumple el

principio de las transferencias». Cuando a es igual a dos, se cumple el principio de las transferencias, pero se ponderan de igual forma las transferencias independientemente de los niveles de renta entre los que se produzcan. Ello quiere decir que no son más valoradas las transferencias entre los más pobres que entre los que tienen una renta intermedia, por ejemplo. Valores superiores a 2 del parámetro a implican que se otorga mayor peso a las transferencias que se producen en los niveles elevados de renta. Para niveles inferiores a 2 (en particular para el índice de Theil) ocurre lo contrario, es decir, que se valoran más las transferencias que se producen en los niveles más bajos de renta, lo que se denomina «Principio de las transferencias decrecientes de renta» [9].

EJERCICIO 12: Obtener el índice de Theil (GE(1)) de la distribución de renta inicial, sin corregir por escalas de equivalencia, y obtener la descomposición en función de los colectivos separados por número de miembros en el hogar.

SOLUCIÓN: La solución se presenta en la subhoja Theil de la hoja de cálculo de ejercicios. En primer lugar obtenemos el índice de Theil para la población en su conjunto, sin desagregar por ninguna característica distintos colectivos. Para ello en la columna D se realizan los cálculos previos del cociente de cada renta con respecto a la media de la distribución total, calculada en la casilla B103 (=B3/$B$103, y se copia hacia abajo para las 100 observaciones de que consta la muestra). En la columna E se obtiene el logaritmo neperiano de la columna anterior (=LN(D3)), para multiplicarlas en la columna F (=D3*E3). Como puede apreciarse. Los valores de la columna D son menores que la unidad para las rentas inferiores a la media y mayores que uno en caso contrario, lo que implica valores negativos o positivos si nos referimos a logaritmos neperianos, obtenidos en la E. Una vez calculado el producto de la columna F se suma para todas las observaciones y se divide entre las observaciones que hay, resultando en la casilla F105(=F103/A102) un índice de Theil de 0,4199.

La segunda parte del ejercicio requiere más pasos, pero el procedimiento es el mismo. Para obtener las rentas por colectivos dependiendo del número de miembros que constan en el hgar, lo más sencillo es ordenar los datos (lo que se presenta en las columnas H, I y J) con la precaución de ampliar la selección. Es decir: ordenamos por número de miembros pero ampliamos el rango de la selección a la renta, para que cada renta vaya emparejada con su número de miembros correspondiente. Una vez ordenado, se separan los colectivos y las rentas se presentan en las colum-

[9] Una explicación más completa de estos conceptos precisaría la consideración de funciones de bienestar social, materia que no se ha querido incorporar en este capítulo de nivel elemental. Para una revisión detallada del tema, puede consultarse Salas (2001).

nas M, R W y AB para los colectivos de 1, 2, 3 y 4 miembros respectivamente. En las columnas contiguas a la renta se obtienen como para el caso general los cocientes con respecto a la renta media (columnas N, S, X y AC) y los correspondientes logaritmos en las columnas (O, T, Y y AD). Por último se obtiene el producto en P, U, Z y AE, que se suma en las casillas marcadas como suma en color rosa. Es importante cuando se calcula la descomposición por subgrupos ser muy cuidadosos acerca del número de miembros que compone cada grupo al calcular las medias subgrupales, lo que será también necesario considerar al calcular la fracción de renta de la que disfruta cada colectivo. Para evitar confusiones, el número de miembros de cada colectivo se ha resaltado en color rojo.

Los índices de Theil de cada uno de los cuatro grupos se presenta en las casillas P27, U27, Z32 y AE34. En el rectángulo encuadrado M31:R36 se presentan todos los valores necesarios para calcular el Theil a partir de la información de subgrupos: en la columna N, los cuatro valores relativos al peso de cada uno de los colectivos respecto del total en cuanto a renta se refiere, lo que implica dividir la renta total de cada grupo por la renta total de la población (por ejemplo, M26*M25/B103, para el primer colectivo). En la columna Q se presentan los productos de dichos pesos por los Theil obtenidos por subgrupos (que se suman en Q36) y en la columna R, se calculan los productos de los pesos por los logaritmos correspondientes (que se agregan en R36). La suma de estos dos valores constituye el índice de Theil, obtenido en la casilla Q37, con idéntico valor al calculado a partir de las observaciones totales, 0,4199. Para comprobar si el valor es el mismo, se obtiene la diferencia con el primer valor en la casilla Q38.

De la observación de los datos recuadrados en la tabla en el rango M31:R36 se puede deducir que el colectivo de 1 y 2 miembros es el que contribuye en mayor medida a la desigualdad total, porque si bien las proporciones en términos de población y de renta (que aunque no son coincidentes evidentemente, pero son todas cercanas a la cuarta parte) son bastante igualitarias, el valor de los índices de Theil de las familias de 1 y 2 miembros es muy elevado en comparación a los otros dos colectivos, que contribuyen con un efecto muchísimo menor. En definitiva, la desigualdad del colectivo de nuestro ejercicio viene explicada en gran parte por la existente en las familias de menos miembros.

6.5. EL BIENESTAR SOCIAL Y LA DESIGUALDAD: ÍNDICE DE ATKINSON

Hasta ahora, y en aras de evitar complicaciones mayores, hemos eludido la consideración de funciones de bienestar social. No es nuestro interés realizar un análisis exhaustivo de tales funciones, pero es necesario referirnos

a algunas cuestiones básicas para presentar el último de los índices de desigualdad, el índice de Atkinson. Este índice, al igual que el de Theil se puede presentar como una familia de índices, ya que hay tantos como valores se le quiera dar al parámetro de aversión a la desigualdad. Esta mayor o «preocupación» por la desigualdad es lo que hace necesario introducir el concepto de función de bienestar social. Al igual que no todos los individuos obtienen la misma satisfacción consumiendo los mismos bienes o disfrutando de la misma renta, no todas las sociedades están igualmente satisfechas en función de las circunstancias de los miembros que la forman. Por ello los mapas de curvas de indiferencia varían entre individuos, e igualmente las funciones que miden el bienestar de una sociedad (funciones de bienestar social) son diferentes dependiendo de las opiniones de la sociedad y el momento del tiempo. Por ejemplo, ante una misma distribución de renta, una sociedad muy preocupada por la desigualdad puede considerar que el bienestar del colectivo es pequeño porque preferiría una situación más igualitaria. Pero otra sociedad menos preocupada por la desigualdad, puede derivar de la misma distribución un bienestar mayor porque la desigualdad le preocupa menos. Nos hallamos ante una cuestión totalmente subjetiva, ya el estar mejor o peor dependerá de los que importe la desigualdad para cada sociedad.

El índice de Atkinson se puede calcular para cualquier función de bienestar social que se desee, si bien al proponer su índice, Atkinson lo acompañó de una función que cumpliera las características que se admiten como deseables para este tipo de funciones. ¿Cuáles son estas características? En principio, y sin consideraciones matemáticas rigurosas, se pueden resumir en tres:

1) Que al mejorar un individuo de la sociedad sin que ninguno empeore, la función de bienestar social arroje un valor superior. Ello implicaría que si aumenta l renta de cualquiera de los individuos de la sociedad —ricos o pobres— la sociedad en su conjunto debe estar mejor. Esta característica no implica más que un criterio de eficiencia Paretiano.

2) Que no importe quiénes sean los miembros de la sociedad que disfrutan de cada nivel de renta, que es lo mismo que exigir anonimato, para no juzgar mejor una situación u otra dependiendo de quiénes sean los que disfrutan de cada nivel de renta. Con ello se pretende evitar discriminaciones de cualquier tipo

3) Que si la renta se distribuye de forma más desigual, el bienestar empeore. Ello implica «aversión a la desigualdad» y quiere decir que la misma renta media satisface más a la sociedad cuanto más igualitariamente se distribuye. Esto es una cuestión de grado, y existen sociedades más aversas a la desigualdad y menos.

Este último punto es clave en la comprensión de los parámetros de aversión a la desigualdad que se incorporan en algunos índices de desigualdad (como Theil o Atkinson) y son la clave para introducir esa subjetividad o preocupación por la desigualdad. La incorporación de los parámetros de aversión a la desigualdad permiten además considerar circunstancias tanto intermedias como extremas: por ejemplo el hecho de que la desigualdad no preocupe en absoluto (solamente importa el tamaño del pastel, pero no cómo está repartido) se modeliza con aversión nula a la desigualdad. El hecho de que la preocupación por la desigualdad sea lo único importante (pasteles más grandes no satisfacen más a la sociedad si los trozos no son idénticos) también se puede incorporar en el cálculo de índices mediante una parámetro de aversión a la desigualdad máximo.

Hechas estas consideraciones previas, pasamos a mostrar la función de bienestar social y utilidad que propuso Atkinson para el cálculo de su índice de desigualdad.

$$U^{i}_{Atkinson} = a + b\ln(x_i) \qquad si\ \alpha = 1$$

$$U^{i}_{Atkinson} = a + b\frac{x_i^{1-\alpha}}{1-\alpha} \qquad si\ \alpha \neq 1$$

$$W_{Atkinson} = \frac{\sum_{i=1}^{N} U_i}{N}$$

La función de bienestar W propuesta por Atkinson es simplemente la utilidad media de todos los miembros de la sociedad, a partir de una utilidad individual que se calcula como la suma de un parámetro *a* más *b* veces la renta elevada a la unidad menos el parámetro de aversión a la desigualdad (a) y dividida entre esa misma diferencia. Para evitar indeterminaciones, se define como a más *b* veces el logaritmo de la renta si el parámetro de aversión a la desigualdad es nulo. Los parámetros *a* y *b* pueden escogerse por conveniencia del investigador, por lo que por simplificar, tomaremos *a* nulo y *b* unitario. En tal caso, veamos qué ocurre en un caso extremo, por ejemplo, si la sociedad no se preocupa en absoluto por la desigualdad, no le preocupa o no le importa en absoluto. En tal caso, a sería nulo, por tanto la utilidad de cada individuo sería coincidente con su renta, y el bienestar de la sociedad sería igual a la renta media. Esta situación viola uno de los principios que hemos marcado como deseables para las funciones de bienestar social, ya que no hay preocupación por la desigualdad: si el bienestar social es igual a la media, da lo mismo que todos los miembros de la sociedad cuenten con la misma renta (la media) que la situación en que todos tienen renta nula excepto un miembro que lo posee todo.

Aún así, es interesante ver cómo los extremos quedan recogidos en la expresión genérica de bienestar propuesta por Atkinson.

A medida que considerásemos valores superiores del parámetro de aversión a la desigualdad, la preocupación por la misma sería mayor, lo que quiere decir que la misma distribución satisface menos a sociedad es más aversas a la desigualdad, o dicho de otro modo, el valor del bienestar social es menor cuanto mayor es el parámetro a para la misma distribución de renta.

Otro concepto que es necesario definir antes de presentar el índice de Atkinson es el de «**renta equivalente igualitariamente distribuida**» que denotaremos por x_d. Esta renta sería aquella cantidad de renta que repartida de forma igualitaria a todos los miembros de la sociedad proporcionaría el mismo bienestar que la distribución actual. Esta renta es menor que la renta media, y es la clave para entender el índice de Atkinson.

Pensemos en términos intuitivos: la sociedad se contenta tanto más cuanto mayor es el pastel con el que cuenta y (si le preocupa la desigualdad) cuanto más igualitarias son las porciones en que se reparte. Por ello, para un tamaño de pastel determinado, el bienestar o satisfacción de la sociedad serían máximos si todos los miembros contasen con la misma porción. En las distribuciones reales, las porciones difieren, y por ello el bienestar no es máximo para un tamaño de pastel determinado, pero podríamos buscar aquella porción de pastel igual para todos los miembros de la sociedad que repartida a todos los miembros lograse el mismo bienestar (que no es el máximo) que con las porciones desiguales. Dicho de otra forma: es posible lograr el mismo bienestar con un pastel más pequeño pero mejor repartido que con uno más grande pero con porciones desiguales. Entonces, ¿Cuánto más pequeño puede ser el pastel si hacemos más iguales los trozos? Eso el lo que trata de captar la renta equivalente igualitariamente distribuida.

Una vez entendido este concepto, la comprensión del índice de Atkinson es inmediata:

$$I_{Atkinson} = A = 1 - \frac{x_d}{\mu}$$

El índice de Atkinson (A) mide la fracción de renta que puede ser sacrificada sin pérdida de bienestar social si la renta fuese distribuida igualitariamente.

De nuevo, pensemos en una situación extrema: si x_d fuese igual a la renta media, querría decir que no se puede sacrificar ninguna renta para conseguir mayor bienestar o A=0, y ello puede ocurrir por dos razones:

a) La distribución ya es igualitaria, por tanto x_d coincide con la renta media. En este caso se trata de una razón objetiva: no existe desigualdad.

b) La distribución no es igualitaria, pero ello no es relevante. Aunque existe desigualdad, no preocupa a la sociedad, que no está dispuesta a sacrificar ninguna porción del pastel por hacerlo más igualitario. Se trata ahora de una cuestión subjetiva: existe desigualdad (y así sería captado por cualquier índice o medida de dispersión de la renta) pero el índice no lo capta porque la aversión a la desigualdad es nula.

El otro extremo lo constituiría aquella situación en que la renta x_d fuese nula. Esto significaría que no hace falta nada de renta para satisfacer a la sociedad en la misma media que ya se hace: si todos tuvieran lo mismo aunque fuese nada, la sociedad sería igual de feliz que con la distribución desigualitaria existente. Ello conlleva que el índice A toma su valor máximo, igual a la unidad, y puede ocurrir en dos casos:

a) La desigualdad es máxima, hay un solo miembro de la sociedad que lo tiene todo, por lo que la renta x_d sería nula: la renta ya es igualitaria para todos menos para uno, y la renta que repartida a todos igual satisfaría igual que no tener nada sigue siendo no tener nada. Se trata en este caso de un hecho objetivo: existe una desigualdad extrema.

b) La preocupación por la desigualdad es máxima, con lo cual se viola incluso el principio de eficiencia Paretiana, y es mejor renunciar a renta con tal de lograr mayor igualdad. Estamos ahora ante una situación subjetiva: existe un grado de desigualdad que no tiene porqué ser el máximo, pero la sociedad lo juzga intolerable.

Así, el índice de Atkinson en situaciones no extremas incorpora elementos tanto objetivos -existe desigualdad- como subjetivos —la desigualdad existente se juzga como más o menos importante—.

EJERCICIO 13: Calcular el índice de Atkinson para la distribución de renta inicial existente sin corregir por escalas de equivalencia. Suponga que la función de utilidad individual y social son las propuestas por Atkinson. Considere qué ocurre cuando se modifica el parámetro de aversión a la desigualdad.

SOLUCIÓN: Se obtiene en la subhoja Atkinson. Lo primero que hay que obtener es la utilidad de cada una de las familias, por lo que es necesario fijar los parámetros que la definen, la aversión a la desigualdad, a, en la casilla E1, y *a* que se ha fijado como 0 y *b* como 1 (aunque se pueden modificar a gusto del usuario) en las casillas E2 y E3 respectivamente. La

utilidad se obtiene para cada observación en la columna C, teniendo la precaución de calcular la utilidad de forma diferente dependiendo de que la aversión a la desigualdad tome o no valor unitario:

=SI($E$1=1;$E$2+($E$3*LN(B5));$E$2+($E$3*(B5^(1-$E$1))/(1-$E$1)))

Lo que responde a la expresión:

Si a=1 $U = a + b\ln(x)$

En otro caso, $U = a + b\frac{x^{1-\alpha}}{1-\alpha}$

Una vez obtenida la utilidad para cada observación, el bienestar social se obtiene como su media aritmética en la casilla E4= =SUMA(C5:C105)/A104.

La renta equivalente igualitariamente distribuida de la casilla E5 se obtiene de despejar la expresión del bienestar que se alcanzaría si todos tuviesen la misma renta igualdo al valor del bienestar que se ha obtenido en la casilla E4, es decir, con la distribución realmente existente: E5=SI(E1=1; EXP((E4-E2)/E3);((1-E1)*(E4-E2)/E3)^(1/(1-E1)))

Si a=1 $x_d = e^{\frac{\bar{W}-a}{b}}$

En otro caso, $x_d = \sqrt[1-\alpha]{\frac{(1-\alpha)(\bar{W}-a)}{b}}$

La media de la renta, necesaria para obtener el índice de Atkinson se presenta en E6 y finalmente, E7 calcula el índice de Atkinson como =1-E5/E6.

Como la hoja se ha construido en términos genéricos, se puede partir de un valor determinado de aversión a la desigualdad e ir comprobando lo que ocurre cuando ésta se modifica. A partir de una situación sin aversión a la desigualdad, con a=0, tendríamos que el bienestar social, y la renta equivalente igualitariamente distribuida coinciden con la renta media (7872,34). Ello implica que el índice de Atkinson vale 0, porque a pesar de que existe desigualdad (obviamente la distribución no es igualitaria) la sociedad no lo juzga como relevante, pues la desigualdad no le importa nada, y el bienestar coincide con la renta media: solamente preocupa el tamaño del pastel, pero el bienestar es el mismo se reparta a partes iguales o lo posea todo una sola unidad. Cuando consideramos que la desigualdad es algo relevante (por ejemplo tomando =0,5, el bienestar desciende hasta 158,899 porque ante el mismo tamaño de pastel (la media sigue siendo

7872,34) la renta equivalente igualitariamente distribuida no es ahora coincidente con la media, sino menor (6312,22). Ello significa que con menos renta pero repartida a todas las familias a partes iguales, podríamos lograr el mismo bienestar que se logra ahora con un pastel más grande pero con porciones desiguales, es decir, se podría sacrificar renta en aras de mayor igualdad sin perder bienestar, y por ello el índice de Atkinson es positivo: 0,1982. Si modificamos el parámetro de aversión a la desigualdad de manera que ésta aumente, lograremos cada vez bienestar menor, renta equivalente menor y mayor valor del índice de Atkinson. El valor no es máximo para a=1, como se podría pensar. En este caso el índice valdría «solamente» 0,3949, siendo el valor máximo la unidad, cuando el parámetro de aversión a la desigualdad es infinito. Puede comprobarse que un valor «infinito» no es excesivamente grande en la práctica, ya que con un valor de a=4 el índice toma un valor de 0,99 y los cambios a partir de entonces se producen de manera lenta (de forma asintótica) cuando se eleva la aversión a la desigualdad. Con los valores *a*=0 y *b*=1 que se han tomado en este ejercicio, lo que sí ocurre es que a partir de aversiones a la desigualdad superiores a la unidad, los niveles de utilidad de las familias toman valores negativos si bien esto no es preocupante a efectos del índice de Atkinson, puesto que nos interesa la magnitud ordinal de la utilidad, y sigue ocurriendo que rentas más elevadas muestran utilidades superiores. De todas formas, como se ha anticipado anteriormente, esto se puede corregir haciendo que *a* no sea nulo, ya que los parámetros de la hoja se pueden modificar siempre que se mantenga la coherencia teórica.

EJERCICIO 14: Construir dos distribuciones alternativas sin modificar la media a partir de la que se propone a continuación de manera que se presenten casos de desigualdad extrema: una sola unidad lo posee todo, o todos cuentan con la renta media. Obtener los valores del índice de Atkinson para diferentes grados de aversión a la desigualdad con la función de bienestar de Atkinson con parámetros *a*=0 y *b*=1.

Distribución propuesta:

Unidad	*Renta*
1	10
2	15
3	40
4	55
5	60
6	100

SOLUCIÓN: Se aporta en las columnas J a P de la subhoja Atkinson. El objetivo de este ejercicio no es aprender a calcular el índice de Atkinson (como se explicó en la solución del ejercicio 13) sino el saber interpretar los resultados que se obtienen en diferentes casos. Para ello hemos forzado a que existan modificaciones objetivas, es decir, que las distribuciones de renta que se analicen sean distintas. Ello se combinará con modificaciones subjetivas, es decir, con la consideración de diferentes aversiones a la desigualdad.

La aversión a la desigualdad se puede modificar en la casilla M1. Comencemos suponiendo una aversión a la desigualdad nula y veamos qué ocurre con los índices de Atkinson en la situación de distribución desigualitaria no extrema, de igualdad total y de desigualdad máxima. En este caso, los valores del índice de Atkinson (celdas P7, P17 y P27 respectivamente) son idénticos y nulos, ya que sea cual sea la distribución existente, no se juzga el reparto, sino el tamaño del pastel (renta media) que es siempre la misma.

Continuemos asumiendo que a=0,5. En tal caso, para la misma aversión a la desigualdad tiene que ocurrir que cuando la distribución es igualitaria, el índice de Atkinson se anule, pero tome valores positivos cuando la igualdad no es perfecta (y mayor en el caso en que una sola unidad lo tiene todo). Eso es lo que efectivamente ocurre. Al mismo tiempo, el bienestar decrece en las distribuciones más desigualitarias, así como la renta equivalente igualitariamente distribuida.

En este caso en que a=0 y b=1 no tiene sentido asumir aversiones a la desigualdad superiores a la unidad, ya que la utilidad y el bienestar serían negativos. Pero incluso bajo estos condicionantes podemos comprobar que el valor máximo del índice de Atkinson surge cuando los factores objetivos y subjetivos confluyen en la peor dirección: la distribución es la más desigualitaria posible (una sola unidad lo posee todo) y la desigualdad se juzga importantísima (a es máximo [10]). En ese caso, el valor del índice de Atkinson es la unidad.

El siguiente cuadro resumen puede ayudar al lector a comprobar la evolución tanto del índice de Atkinson como de la renta equivalente en las tres distribuciones propuestas:

[10] El valor de a sería 1, pero tomamos 0,9999999 porque dadas las pocas observaciones en el ejercicio, es más fácil obtener así los valores de Atkinson=1.

CUADRO 5.1. *Resultados del índice de Atkinson para distinta aversión a la desigualdad ante tres distribuciones diferentes*

a	*Distribución igualitaria*		*Reparto desigual*		*Una sola unidad lo posee todo*	
	x_d	A	x_d	A	x_d	A
0	46,6	0	46,6	0	46,6	0
0,3	46,6	0	43,5	0,06	21,6	0,53
0,5	46,6	0	41,2	0,12	7,8	0,83
1	46,6	0	35,4	0,24	2,6	1

6.6. CONSIDERACIONES FINALES

Tras la revisión de este sexto capítulo el lector es capaz de caracterizar de múltiple formas la desigualdad existente en una distribución de renta y calcular los índices más utilizados para este mismo fin. No por haber utilizado un enfoque sencillo se ha perdido en rigurosidad ni han dejado de incluirse cuestiones fundamentales. Una vez que se conocen estas herramientas, el lector puede abordar sin excesiva dificultad el estudio de la contribución al efecto redistributivo de las estructuras impositivas que gravan la renta. Básicamente se trata de comparar situaciones pre-reforma y post-reforma (o anteriores y posteriores al pago de impuestos), y profundizar en el manejo de curvas de Lorenz e índices de progresividad. Los contenidos del presente capítulo capacitan al lector para seguir avanzando en el estudio de la distribución de la renta y análisis de reformas fiscales, pero por sí solos permiten analizar la relevante cuestión de la desigualdad de forma cuantitativa y sustentada en la maximización del bienestar.

Con el fin de facilitar que el lector aprenda por su cuenta y profundice en el estudio de la distribución y desigualdad de la renta, se aportan algunos sitios web que pueden resultar interesantes. La mayor parte del material se encuentra en inglés. Muchos organismos ofrecen material para realizar cursos, o bien permiten el acceso libre a documentos de trabajo, y publican información acerca de eventos de interés.

— El Banco Mundial realiza conferencias repetidamente en relación con cuestiones de desigualdad. Los trabajos que allí se presentan se ponen a disposición de los interesados. Para los eventos de los años 2002, 2003 y 2006, se puede encontrar la documentación en la página:

http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTPROGRAMS/EXTPOVRES/EXTDECINEQ/0,,contentMDK:20553503~menuPK:1359622~pagePK:64168445~piPK:64168309~theSitePK:1149316,00.html

— También desde la página Web del Banco Mundial se ofrece material para la realización de un curso sobre desigualdad con una excelente selección de artículos relevantes para las materias referentes a de quién se mide la desigualdad, cómo medirla, cuáles son los determinantes y cómo influyen las variables macroeconómicas en la desigualdad. El material se puede descargar en: http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTPROGRAMS/EXTPOVRES/EXTDECINEQ/0,,contentMDK:20553458~menuPK:1359587~pagePK:64168445~piPK:64168309~theSitePK:1149316,00.html

— Otro curso de desigualdad ofrecido desde la página del Banco Mundial es el desarrollado por la Universidad de Maryland, con el objetivo de mostrar la desigualdad global de los ciudadanos en el mundo, y de trazar las consecuencias éticas y políticas de la desigualdad hallada. Se puede seguir en sus tres secciones en: http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTPROGRAMS/EXTPOVRES/EXTDECINEQ/0,,contentMDK:21186720~menuPK:1359587~pagePK:64168445~piPK:64168309~theSitePK:1149316,00.html

— La Sociedad para el estudio de la desigualdad económica incluye un apartado de documentos de trabajo relativos a desigualdad. Su dirección es: http://www.ecineq.org/

— El Instituto de Estudios Fiscales Británico ofrece una página web con publicaciones relativas a desigualdad, pobreza y bienestar en la dirección: http://www.ifs.org.uk/publications.php?heading_id=8

BIBLIOGRAFÍA

Buhmann, B.; Rainwater, L.; Schmaus, G. y Smeeding, T.M. (1988). «Equivalence scales, well-being, inequality and poverty: sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database». *Review of Income and Wealth, 34, pp. 115-142*.

Citro C., y R. T. Michael, ed. (1995). *Measuring Poverty. A New Approach*. National Academy Press, Washington, D.C.

Lambert, P. (1996) *La distribución y redistribución de la renta: un análisis matemático*. 2ª ed. Instituto de Estudios Fiscales.

Mancero, X. (2001). *Escalas de equivalencia: reseña de conceptos y métodos.* Nº 8 Serie estudios estadísticos y prospectivos. CEPAL. Santiago de Chile. http://www.eclac.org/publicaciones/xml/9/6569/lcl1492e.pdf

Ruiz-Huerta Carbonell, J. ed (2005). *Políticas públicas y distribución de la renta.* Fundación BBVA, Bilbao.

Salas, R. (2001) «La medición de la desigualdad económica». *Papeles de trabajo del Instituto de Estudios Fiscales,* nº 14. Madrid. http://www.ief.es/Publicaciones/PapelesDeTrabajo/pt2001_14.pdf

Silber, J. ed. (1999) *Handbook of income inequality measurement*. Kluwer Academic Publishers.

CAPÍTULO VII
MEDICIÓN DE LA POBREZA

NURIA BADENES PLÁ

7.1. INTRODUCCIÓN

La medición de la pobreza debe abordarse como paso posterior al análisis de la distribución de la renta, por lo que se recomienda al lector que se familiarice con los conceptos que se exponen en el Capítulo 6. Cuando se analiza la pobreza el interés se concentra ya no en la distribución total de la renta sino en un colectivo muy determinado: los más desfavorecidos, o los que se sitúan en la cola baja de la distribución.

La pobreza está asociada en la mente de todos a connotaciones negativas. Pero ¿qué significa exactamente ser pobre? Los gobiernos de los países, el Banco Mundial, el Fondo Monetario Internacional y otros organismos y ONG's necesitan saber quiénes son los pobres, y en que grado de severidad sufren la pobreza para orientar correctamente sus políticas de ayuda. Ser pobre no puede reducirse exclusivamente al hecho de carecer de dinero. Aunque por tratarse de un capítulo introductorio adoptaremos aquí un enfoque unidimensional que resulta más sencillo, en realidad la pobreza es un fenómeno multidimensional, que puede derivar de circunstancias no materiales y ambientales como las catástrofes naturales, los problemas políticos, la crisis de los mercados financieros, o la discriminación, por citar algunas.

Asimismo, las manifestaciones de la pobreza son diversas y terribles todas ellas, relacionadas con la imposibilidad de garantizar el acceso o el consumo mínimo de bienes preferentes, lo que se traduce en hambre, enfermedad, analfabetismo, y/o bajas esperanzas de vida. Estas circunstancias se consideran en el IPH (Indice de pobreza humana) propuesto por las Naciones Unidas en 1997, según el cual la cuarta parte de la población mundial sería pobre —y lo sería más en el África subsahariana y en Asia meridional—. Si se quiere utilizar una medida más sencilla y unidimensional, se fija una línea o umbral de pobreza, nivel por debajo del cual se considera que alguien es pobre. Con el fin de llevar a cabo comparaciones internacionales, se suele fijar esta línea en un dólar diario por persona, suma considerada suficiente para adquirir los productos necesarios para sobrevivir. Actualmente, en el mundo en desarrollo 1.300 millones de personas viven con menos de un dólar diario y cerca de 3.000 millones —casi la mi-

tad de la población mundial—, con menos de dos dólares. Vistos estos datos, queda patente que la pobreza constituye un problema de suficiente envergadura como para dedicarle estudio, y además, que la medición del mismo es clave para poder combatirlo.

Así como el fenómeno de la desigualdad de la renta puede decirse que constituye hoy un tema arraigado tanto en el ámbito académico como investigador y de análisis empírico, la pobreza ha cobrado importancia —al menos en el ámbito académico— de forma relativamente reciente. Ello explica que no abunden los manuales que utilicen un enfoque didáctico para la explicación de la pobreza, y sea necesario acudir a las fuentes originales.

El presente capítulo trata de explicar de forma simplificada y para aquellos que se acercan por primera vez a la medición de la pobreza en qué consiste el fenómeno, cómo medirlo mediante los índices más comunes y cómo construir curvas TIP. Se plantea también la multidimensionalidad del fenómeno. No se abordan otros tópicos puesto que el material presentado solamente pretende dar una visión general a aquellos que estudian la pobreza por primera vez. Para completar la visión teórica, se proponen ejercicios que se pueden resolver en hoja de cálculo y cuya solución se aporta y explica.

El capítulo se organiza como sigue: tras esta introducción, el segundo epígrafe se dedica a la explicación de los indicadores de bienestar habitualmente utilizados, ya que la pobreza se define precisamente como ausencia de bienestar. Se aborda además la delicada cuestión de la elección de la línea de pobreza. El tercer epígrafe presenta una selección de los índices más sencillos y que al mismo tiempo son comúnmente utilizados en el trabajo empírico, explicando cuáles de las propiedades deseables cumple cada uno de ellos. El cuarto apartado explica el concepto de curva TIP, que permite de forma gráfica y sencilla describir las dimensiones de incidencia, intensidad y desigualdad de la pobreza de una distribución. En el quinto epígrafe se aborda la multidimensionalidad del fenómeno de la pobreza, explicando las limitaciones del enfoque unidimensional. El sexto epígrafe concluye y presenta algunas direcciones de páginas web de interés para el estudio de la pobreza.

7. 2. INDICADORES DE BIENESTAR

Es necesario definir y medir la pobreza para poder combatirla, porque si no se define un fenómeno no se puede medir, y si no se mide es como si no existiera. El Banco Mundial define la pobreza como la «falta pronunciada de bienestar». Hay que aclarar entonces lo que es el bienestar, que según la visión convencional vincula el concepto a la disponibilidad de renta suficiente para alcanzar el consumo básico de ciertos bienes.

El primer paso necesario en la medición de la pobreza, es entonces definir un indicador del bienestar. Los candidatos que generalmente se utilizan son la renta y el consumo, pero es preciso matizar algunas cuestiones:

a) Si se pretende analizar la pobreza de los individuos y los datos con los que se cuenta corresponden a hogares, una solución es calcular la renta o el consumo *per capita* del hogar. Pero esto supone hacer supuestos irrealistas, porque no todos los miembros del hogar aportan la misma renta ni consumen lo mismo. Realizar repartos igualitarios entre miembros de un hogar implica que no se computan las diferencias en necesidad de sus miembros ni las economías de escala por vivir en familia y compartir gastos. Esto puede resolverse en parte mediante el uso de escalas de equivalencia, pero su uso es bastante controvertido, pues no existe un acuerdo generalizado acerca de las ponderaciones que deben otorgarse a los miembros de una familia.

b) La medición adecuada de la renta para determinar el bienestar puede no ser fácil de obtener. Por ejemplo, la definición de renta extensiva de Haig y Simons exigiría conocer el consumo y el aumento de la riqueza neta, lo que no siempre está disponible. Tampoco está claro cuál es el período idóneo para el cómputo de la renta, si un mes, un año o toda la vida. Por otro lado, las encuestas no siempre constatan la verdad de la renta pues existen reticencias a declarar, olvidos y otras razones que pueden llevar a la infraestimación de la misma.

c) La cuantificación del bienestar y la pobreza a través del consumo, más adecuada en ciertos países subdesarrollados que la renta, tampoco está exenta de problemas de medición, especialmente en el caso del autoconsumo o en la valoración de bienes duraderos. Por otro lado, puede no existir correspondencia entre el bienestar y el consumo en caso de unidades ricas pero austeras.

Existen otros indicadores de bienestar, como la ingesta calórica por persona y día (que suele fijarse en 2.100 calorías diarias). Puede establecerse un mínimo calórico necesario por debajo del cual se consideraría que una persona es pobre, si bien las circunstancias como el sexo, la edad o la actividad física deberían ser tomadas en cuenta. Otro indicador de bienestar es la proporción de renta total que se destina a alimentos, ya que como observó Engel, a medida que la renta crece lo hace el gasto en alimentos, pero en menor proporción. Puede fijarse entonces una proporción por encima de la cual una unidad sería pobre. Otros indicadores colectivos, no referidos a individuos u hogares que pueden utilizarse son aquellos que

captan resultados como la esperanza de vida, la tasa de mortalidad infantil, o el porcentaje de escolarizados. Estos indicadores no se plantean como sustitutos de la renta o el consumo, sino más bien como complementarios en la descripción y medición de un fenómeno multidimensional.

7.2.1. Elección de la línea de pobreza

Las líneas de pobreza marcan el límite que separa los pobres de los no pobres. Es decir las líneas de pobreza marcan el consumo o la renta mínimos necesarios para que una unidad escape de la pobreza. Se pueden fijar líneas alternativas para capturar quiénes son pobres (línea de pobreza), quiénes son extremadamente pobres (línea de pobreza extrema), o quienes no consiguen el alimento básico (línea de pobreza alimenticia).

Una línea de pobreza z_i se puede definir como la cantidad necesaria de gasto o consumo para que la unidad i, con características demográficas x, y dados los precios p, sea capaz de disfrutar del nivel de utilidad u_z.

$$z_i = e(p, x, u_z)$$

Según esta definición, existiría una línea de pobreza diferente para cada unidad, si bien en la práctica, lo que se tiene en cuenta es una agregación de las circunstancias y se utiliza la misma línea de pobreza para todo un colectivo de unidades.

Las líneas de pobreza pueden cambiar, bien porque se desea incorporar el efecto de la inflación o bien porque a medida que el tiempo pasa y las sociedades mejoran, se elevan los estándares de bienestar y las líneas se sitúan por encima [1]. Pensemos, tras la definición ofrecida de línea de pobreza, qué ocurriría si los precios se duplicaran: sería mucho más difícil adquirir bienes si la renta no se hubiese duplicado igualmente, por lo que dejar la línea de pobreza en el nivel anterior a la subida de los precios implicaría que el cómputo de pobres estaría infraestimado. Del mismo modo, si ciertas comodidades pasan a considerarse indispensables en una sociedad (por ejemplo la calefacción, el agua corriente o la ducha en el interior de la vivienda) deberá elevarse la línea de pobreza, ya que la renta necesaria para adquirir elementos extra o el valor de este consumo nuevo implican un z_i mayor.

[1] Un estudio de Ravallion, Datt y van de Walle (1991) referido a 36 países revela que la elasticidad de la línea de pobreza ante cambios en el consumo *per cápita* es de alrededor del 0.7. Ello significa que si el consumo *per capita* se eleva en un 1%, la línea de pobreza lo debería hacer en un 0,7%. Pero estos resultados se refieren a la media de países, ya que los países de más renta presentan elasticidades cercanas a al unidad, mientras que los de menor renta, cuentan con elasticidades casi nulas.

Esta idea nos da pie a la explicación de la diferencia entre **pobreza relativa y pobreza absoluta**. La medición de la pobreza relativa implica establecer una línea de pobreza según los valores que se derivan de la distribución de población analizada, por ejemplo, es pobre todo aquel que cuenta con menos de la mitad de la renta media de la distribución. Una visión relativa implica que siempre habrá pobres en una distribución. La alternativa es la medición de la pobreza en términos absolutos, fijando la línea de pobreza independientemente de los valores de la distribución de la población analizada, por ejemplo, definir como pobres todas las unidades que viven con menos de un dólar diario. Normalmente, las líneas de pobreza absolutas se fijan de manera que no haya de revisarse la definición de pobreza, lo cual es esencial cuando se plantean objetivos de lucha contra la pobreza que abarcan más de un año.

La medición de la pobreza según cada uno de estos enfoques ofrece resultados muy diferentes. Si se adopta un enfoque relativo, el número de pobres puede ser muy parecido en España y en Indonesia. Pero si se utiliza la línea de un dólar diario, posiblemente no existan apenas pobres en España, mientras que sí los habrá en Indonesia. Cuando se quieren realizar comparaciones entre países, debe adoptarse un enfoque absoluto para que los resultados tengan sentido.

Cuando se mide la pobreza en **términos absolutos**, la elección de la línea de pobreza, da lugar a ciertas controversias. En primer lugar existe un problema de identificación del nivel de utilidad: ¿cuál debe ser el nivel de utilidad que se pretende alcanzar y para el cual es necesario determinar la renta o el gasto mínimos?, o dicho de otra forma ¿cuál es el nivel de utilidad alcanzado cuando una unidad se sitúa exactamente en la línea de pobreza?. En segundo lugar, incluso si determinamos ese nivel de utilidad, existe un problema de referencia: ¿cuál es el nivel de renta o gasto necesarios para alcanzarlo?. Esta cuestión surge del hecho de que las unidades, —normalmente hogares— difieren en su composición y no cuesta lo mismo alcanzar un nivel determinado de utilidad si se tiene un hijo que si se tienen cinco.

Una alternativa de diseño de líneas de pobreza absolutas es lo que se conoce como **líneas de pobreza objetivas**, que se basan en determinar el nivel que permite a las unidades adquirir ciertas capacidades entre las que se incluyen una vida sana y activa y con participación en la sociedad. Ello se concreta en dos métodos conocidos como *coste de las necesidades básicas* e *ingesta de energía alimenticia*. El método del coste de las necesidades básicas estima el coste de consumir bienes alimenticios (suponiendo una dieta de 2.100 calorías al día) y no alimenticios básicos (espacio en el hogar, electricidad, etc). Este método requiere conocer el precio de los bienes consumidos por los pobres, lo que no siempre es fácil. Como alternativa se pue-

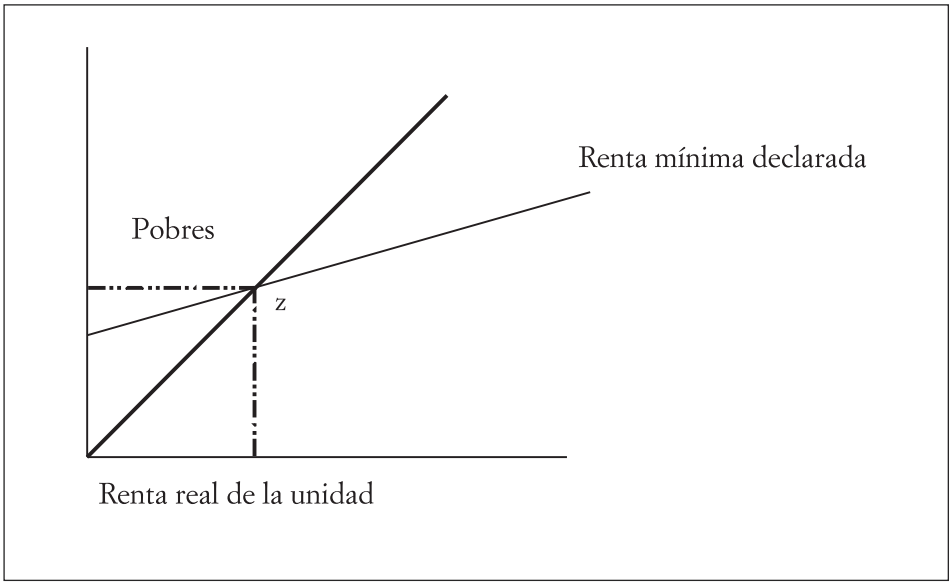


GRÁFICO 2.1.1. *Determinación de la línea subjetiva de pobreza*

de utilizar el método de ingesta de energía alimenticia que relaciona consumo de calorías con renta, escoge un nivel mínimo calórico y determina la renta necesaria para alcanzarlo sin necesidad de conocer los precios. El método adolece de varios problemas, por ejemplo, la dieta varía en el ámbito urbano y rural, según la actividad realizada, el coste de los alimentos es diferente por zonas, y la elección de los mismos puede condicionarse por la publicidad. Por otra parte, la renta necesaria para alcanzar 2.100 calorías diarias puede variar en el tiempo, sobre todo si los precios de los bienes alimenticios se modifican mucho en relación a los no alimenticios y las unidades cambian su consumo de alimentos.

Otra alternativa de diseño de líneas de pobreza absolutas es el **enfoque subjetivo**, que se basa en determinar el nivel mínimo de renta que para una unidad es necesario, y por debajo del cual no podría adquirir lo básico. Si representamos la renta mínima declarada para cada nivel de renta verdadero junto a la bisectriz, para todos los que la renta real sea inferior a la mínima ocurrirá que son pobres, y los que su renta supere la mínima, no serán pobres, determinando el cruce de la renta mínima y la bisectriz el mínimo que marca la línea subjetiva de pobreza.

Las líneas de pobreza subjetivas también cuentan con desventajas. El trabajo empírico pone de manifiesto que las rentas mínimas declaradas son elevadas, no difieren de forma significativa entre los pobres y quienes no lo son,

son prácticamente el doble en las zonas urbanas que en las rurales, y aumentan mucho en el tiempo (ente los ochenta y los 90 se elevaron entorno al 60%).

7.3. LOS ÍNDICES DE POBREZA

Los índices que habitualmente se usan en los trabajos que cuantifican la pobreza como deficiencia de renta, necesitan en primer lugar establecer un nivel a partir del cual se considera que una unidad es pobre, es decir, cuál es el umbral o la línea de pobreza. Ello determinará que todas las unidades con renta inferior a tal nivel, son pobres. Esta es una definición demasiado escueta de pobreza si aceptamos la multidimensionalidad del fenómeno, pero a cambio es muy sencilla, y operativa, porque permite realizar comparaciones de forma inmediata. Como se ha expuesto en el apartado anterior, la fijación del umbral de pobreza cuenta con la desventaja de la arbitrariedad y es necesario que este umbral tenga en cuenta la sociedad que se esté considerando, ya que no es lo mismo contar con un determinado nivel de renta en Marruecos o en Noruega. Para evitar estos problemas se puede fijar el umbral ligado a la distribución (línea de pobreza relativa en lugar de absoluta). Además de las críticas ligadas a la elección del umbral de pobreza, los índices que vamos a exponer son criticados por el hecho de que solamente tienen en cuenta a la población que es clasificada como pobre ignorando lo que ocurre por encima del umbral, o bien no tienen en consideración la intensidad del fenómeno, o dicho de otra forma, cuán pobres son los pobres. Para entender el sentido de tales críticas es necesario exponer las características que se consideraría deseable que cumpliesen los índices y detallar cuáles se cumplen y cuáles no en cada caso.

7.3.1. Propiedades deseables

Una vez que se ha establecido un umbral que separe los pobres de los no pobres, es posible computar cuántos pobres hay en una distribución, cómo de intenso es el fenómeno que sufren o la desigualdad dentro del colectivo, mediante el uso de los índices de pobreza. Previamente a la exposición de los mismos, se presentan nueve características que sería deseable que los índices cumpliesen. Suponemos para su explicación que la única variable sobre la que se mide la pobreza es la renta.

1) Dominio: Si se producen cambios en la renta de los ricos que no afectan a la renta de los pobres, el valor de la pobreza no varía. Es decir, el nivel de pobreza es independiente de la renta de los no-pobres.

2) Monotonicidad: Si la renta de los pobres se reduce, el índice de pobreza debe aumentar de valor

3) Simetría o Anonimato: Si se produce una permutación entre pobres, es decir, dos individuos pobres intercambian sus posiciones, el valor del índice de pobreza no varía.

4) Independencia a la replicación: Si varias subpoblaciones se agregan, el índice de pobreza no varía.

5) Sensibilidad a las transferencias: Si se produce una transferencia entre individuos pobres del más pobre al menos pobre, el valor de la pobreza aumenta

6) Sensiblidad decreciente a las transferencias: Si se produce una transferencia entre un pobre hacia otro menos pobre, comparando esta situación con otra en la que la distancia entre donante y receptor sea la misma pero el donante sea más pobre, ocurre que el índice de pobreza aumenta más en el segundo caso que en el primero.

7) Monotonicidad subgrupal: Si la pobreza en un subgrupo aumenta, permaneciendo constante la de los demás grupos, el valor de la pobreza aumenta

8) Descomposición aditiva: La pobreza global se puede descomponer en suma ponderada por las proporciones sobre el total de cada grupo de la pobreza de subgrupos.

9) Crecimiento de los no pobres: Si la distribución cambia porque aparece una unidad que no es pobre, la pobreza disminuye.

Mostramos a continuación de forma resumida, entre los índices básicos que se van a explicar, qué propiedades de las presentadas como deseables cumple cada uno de ellos. El lector deberá volver a este resumen de propiedades una vez que haya comprendido cómo se calculan los distintos índices de pobreza

Además de las propiedades de los índices de pobreza, existen otros condicionantes que pueden dotar o no de robustez a los índices de pobreza, no ya por la bondad del instrumento de medida sino por la calidad de los datos sobre los que se mide, como son:

a) Errores de muestreo: como la mayor parte de las veces no se cuenta con la información de la población total sino de una muestra, pueden existir errores de muestreo que hagan que la medición de la pobreza realizada no coincida con la verdadera magnitud. Ello se puede corregir aportando las desviaciones e intervalos de confianza de los índices.

Propiedades que cumplen los índices de pobreza básicos

Índice	*Dominio*	*Monot.*	*Simetría*	*Indepen. Pobla.*	*Transf.*	*Transf. Decrec.*	*Monot. Subgroup.*	*Descomp.*	*Crec. no pobres*
H	Sí	No	Sí	Sí	No	No	No	Sí	Sí
PG	Sí	Sí	Sí	Sí	No	No	Sí	Sí	No
I	Sí	Sí	Sí	Sí	No	No	Sí	No	No
HI	Sí	Sí	Sí	Sí	No	No	Sí	Sí	Sí
FGT	Si a>=2	Sí	Sí	Sí	Sí	Si a>2	Sí	Sí	Sí

b) Errores de medida de los datos por la infraestimación o infradeclaración de la renta y el gasto. Este es un problema común, y una infradeclaración del 5% puede elevar en un 10% algunos índices de pobreza (Véase World Bank).

c) La elección de la línea de pobreza lleva aparejado un nivel de pobreza, y en ocasiones esta elección puede ser arbitratria [2].

d) El uso de escalas de equivalencia puede condicionar los resultados. No existe un acuerdo unánime acerca de cuáles son las ponderaciones que deben utilizarse, y por ello el uso de las mismas no está demasiado extendido.

7.3.2. Cuantificación de la pobreza: índices unidimensionales

7.3.2.1. Headcount ratio

El primero de los índices que exponemos se denomina *headcount ratio* (*H*) y mide la proporción de unidades pobres respecto de la población total. Denominando *q* al número total de unidades pobres y *N* a la población total, se tiene que:

$$H=\frac{q}{N} \qquad [1]$$

Este índice no tiene en cuenta ni la intensidad de la pobreza ni la desigualdad entre los pobres, ya que pase lo que pase con las rentas de los po-

[2] Como se explicará en el apartado de curvas TIP, cuando no existe dominancia estocástica de primer orden, la elección de la línea de pobreza es determinante en las conclusiones que se derivan de las comparaciones entre distribuciones alternativas.

bres, no queda reflejado en el índice a no ser que se rebase la línea de pobreza. Por ejemplo, si en una población de 50 unidades hay 25 que no sobrepasan el umbral de pobreza, *H* vale 0,5, indicando que el 50% de la población es pobre. Pero no se cuenta con ninguna información que diga si esos pobres están cerca o lejos del umbral de pobreza, lo cual es muy diferente. Si todos los pobres lo son porque les falta solamente una unidad monetaria marginal para rebasar el umbral, la pobreza es poco intensa, mientras que si todos los pobres cuentan con cero unidades monetarias, el fenómeno es más intenso y esto no queda reflejado en una medida como *H*. Si todos los pobres duplicasen su renta pero aún así continuasen sin rebasar la línea de pobreza (es decir, siendo pobres), *H* no apreciaría ningún cambio de medida de pobreza. O si se produjesen trasvases de renta «desigualitarios» de forma que los más pobres donasen renta a los menos pobres, que están más cércanos al umbral, pero no cambiase el número de pobres, *H* tampoco lo tendrían en cuenta.

EJERCICIO 1: Calcular *H* con la distribución de renta que se adjunta en la hoja de ejercicios correspondiente al capítulo 7, y comprobar su valor si todos los pobres contasen con una renta nula. Para ello considere en primer lugar que la línea de pobreza se establece en el 50% de la renta media, y compruebe cómo se modifica el valor de H para valores inferiores o más elevados de la línea de pobreza.

SOLUCIÓN: Se obtienen los resultados en la subhoja *Headcount ratio* de la hoja de soluciones correspondientes al capítulo 7. Lo primero que hay que hacer es contar con una distribución de renta ordenada, como es el caso. Calculamos la renta media para fijado un porcentaje deseado de la misma, establecer el valor de la línea de pobreza. La media se obtiene en la celda B104 como =SUMA(B3:B102)/A102. En B105 se puede escoger el valor del porcentaje de la renta media que supone la línea de pobreza (expresado en tanto por uno). El ejercicio exige que este valor sea 0,5, dando lugar a una línea de pobreza igual a 2734,92 unidades monetarias. La columna C calcula un indicador para cada una de las observaciones que vale 1 si la unidad es pobre (su renta es inferior a la línea de pobreza) y 0 en caso contrario. Sumando el número de valores 1 contabilizamos la población pobre (C108), y dividiéndola entre la población total, se obtiene el headcount ratio (*H*) en la casilla C109. En este caso se obtiene un 19% de población pobre.

Si toda la población pobre contase con renta nula, el valor de *H* no se vería modificado, ya que este índice no tiene en cuenta la renta de los pobres para medir la pobreza, sino cuántas personas son pobres, y esto no ha variado. Para comprobarlo (y especialmente para facilitar cálculos posteriores), se modifica la renta de los pobres en la columna E de manera que si la unidad es pobre, su renta se sitúa en cero, y en otro caso, se deja

invariable (=SI(C3=1;0;B3)). En la columna G se efectúa de nuevo el recuento de pobres utilizando el indicador 1 para quien es pobre y 0 para quien no lo es (=SI(E3<$B$106;1;0)). La casilla G108 presenta el número de pobres, y la G109 el ratio respecto de la población total, que como se puede comprobar, es idéntico al obtenido previamente (C109).

Tal y como se ha construido *H*, observar su variación ante modificaciones de la línea de pobreza es automático. En la casilla B105 tomamos valores diferentes y automáticamente se producen los cambios en el resultado. Por ejemplo, considerar pobres a aquellos que cuentan con menos de un 25% de la renta media de la distribución implica contabilizar una pobreza extrema, es decir que el valor de los que están tan mal situados desciende respecto a una línea de pobreza de la mitad de la renta media, en concreto, *H* toma valor del 12% en este caso. Si la línea de pobreza se fija en un valor más elevado, digamos que es pobre todo aquel que tiene menos que la media, B105 se fijaría en 1 y habría un 48% de población pobre. En definitiva, líneas de pobreza más bajas implican menor número de pobres (y su peso respecto al total de la población, *H* será también menor), y más elevadas, mayor número de pobres (y mayor *H*). Además, se puede comprobar que el valor de *H* es siempre el mismo computando la renta inicial de los pobres o suponiéndola nula para todos ellos, independientemente del valor que tome la linea de pobreza (C109 y G109 coinciden en todo caso para cualquier valor supuesto de C105).

7.3.2.2. Poverty gap

Otro índice de pobreza sencillo es el llamado *poverty gap (PG)* o brecha de pobreza, que mide la renta que sería necesario otorgar a los pobres para que dejasen de serlo. Siendo x_i cada una de las rentas, se calcula como:

$$PG=\sum_{i=1}^{q} z-x_i \qquad [2]$$

Es decir, que el índice se construye agregando para todos aquellos que son pobres, los déficits de renta hasta la línea de pobreza. Al contrario de lo que ocurre con *H*, *PG* sí que es capaz de medir la intensidad de la pobreza, y arrojaría valores diferentes si todos los pobres lo fuesen pero estuvieran muy cercanos al umbral o si lo fueran porque sufren una carencia mucho mayor.

EJERCICIO 2: Calcular *PG* con la distribución de renta existente, y comprobar su valor si todos los pobres contasen con una renta nula y si todos los pobres contasen con la misma renta que el pobre más rico.

SOLUCIÓN: Los cálculos correspondientes al ejercicio 2 se presentan en la subhoja poverty gap. En primer lugar se calculan los déficits correspondientes a la distribución original de la columna B, de forma que se toma la información relativa a si la unidad considerada es o no pobre (columna C) para ver si se ha de computar o no un déficit. Así, en la columna D se calcula la renta que falta hasta la línea de pobreza en caso de que el indicador de la columna C tome valor igual a 1 (=SI(C3=1;$B$106-B3;0)). Si la unidad no es pobre, entonces no se computa déficit alguno. Para un valor de la línea de pobreza igual al 50% de la renta media, *PG* toma un valor de 33825,99 unidades monetarias, que es la renta que sería necesario transferir a los pobres para acabar con la pobreza. Para ello el reparto no podría realizarse de cualquier manera, ya que habría que repartir a cada uno de los pobres la cantidad que le falta hasta alcanzar la línea de pobreza.

Si consideramos ahora distribuciones alternativas de renta entre los pobres, el índice sí que se modifica aunque se mantenga el número de pobres. Para comprobarlo se propone calcular *PG* cuando todos los pobres cuentan con renta nula. En ese caso, la distribución necesaria para realizar los cálculos es la misma que se propuso en el ejercicio 1, y aparece en la columna G. El indicador de si una unidad es pobre o no, no varía con respecto a la distribución anterior (columnas C y H coincidentes). Pero en este nuevo escenario, el déficit de cada pobre es exactamente la línea de pobreza, ya que a todos les falta tanto como el nivel establecido para separar pobres de no pobres. Solamente en el caso de que todos los pobres contasen con renta nula, *PG* podría obtenerse de forma muy sencilla como el producto del número de pobres por la linea de pobreza (casilla H108), que coincide con el valor calculado de forma ortodoxa en I104. Para llegar a obtener este resultado, se han agregado todos los déficit calculados en la columna I. Como puede comprobarse, para una línea de pobreza del 50% de la renta media, PG vale 51963,45, valor superior al obtenido cuando no todos los pobres contaban con renta nula. Es decir, que este índice sí que es sensible a la intensidad de la pobreza, o considera cuán pobres son los pobres.

Un nuevo escenario sería considerar que todos los pobres cuentan con la renta máxima de entre los pobres. Ello reflejaría una situación de pobreza nada extrema, al contrario que en el caso anterior, ya que con un poco de renta para cada pobre se acabaría con la pobreza. Lo primero que debemos calcular es el valor de la renta que debe tener cada pobre. Para ello se utiliza la columna K como comodín intermedio. Se obtiene aquí el producto de la renta por el indicador de si se es o no pobre (=B3*C3). Así contaremos con ceros en el caso de que la unidad sea no pobre, y con la renta si la unidad es pobre. La renta que nos interesa es el máximo de ese valor, que se obtiene en L104 (=MAX(K3:K102)). Esa es la renta que ad-

judicamos a los pobres, dejando a los no pobres como están. Esta nueva distribución se escribe en la columna L (=SI(C3=1;$L$104;B3)). En la columna M se vuelve a obtener un indicador de si la unidad es o no pobre conforme a la nueva distribución (que obviamente dará los mismos resultados, ya que las modificaciones supuestas afectan a la renta entre los pobres, pero no modifica quiénes son pobres). Por último se calculan los déficits entre la linea de pobreza y la renta para aquellas unidades que son pobres (=SI(M3=1;$B$106-L3;0)). Para el mismo valor de línea de pobreza que en los ejemplos anteriores (50% de la renta media) *PG* toma un valor de «solamente» 10979,71 unidades monetarias, indicando que hace falta menos renta que en los casos anteriores para hacer que los pobres dejen de serlo.

La hoja de cálculo está construida para que se puedan comprobar los resultados con distintos valores de línea de pobreza, siempre expresada como un porcentaje de la media (casilla B105), por lo que el lector puede comprobar que si bien para líneas de pobreza más elevadas los valores de PG siempre aumentan, se mantiene la ordenación obtenida en el ejercicio propuesto:

PG (pobres renta=0) > PG (pobres distribución cualquiera) > PG (pobres renta=máxima entre pobres)

7.3.2.3. Income gap ratio

El siguiente índice que se presenta es el *income gap ratio (I)* que sí tiene en cuenta los déficits de renta de aquellos que son pobres pero a cambio no tiene en cuenta la proporción que representan los pobres en el total de la población. Se calcula agregando todos estos déficits y la suma total se divide entre el producto del número de pobres por la línea de pobreza. Si denominamos x_i a cada una de las rentas y $\mu(q)$ a la renta media de los pobres, podemos expresar *I* de estas dos formas alternativas:

$$I = \frac{\sum_{i=1}^{q} z - x_i}{qz} = 1 - \frac{\mu(q)}{z} \qquad [3]$$

Lo que indica *I* es la proporción entre la renta que habría que dar a los pobres para que dejasen de serlo y la que tendrían los pobres si se situasen sobre el umbral de pobreza. Nótese que para el mismo número total de pobres, si la intensidad de la pobreza es severa, el valor de *I* será mayor que si los pobres tienen una renta cercana a la línea de pobreza.

EJERCICIO 3: Calcular *I* con la distribución de renta existente, y comprobar su valor en los siguientes escenarios alternativos:

a) Todos los pobres tienen renta nula.

b) Todos las observaciones de la distribución inicial duplican su renta.

SOLUCIÓN: Se presentan en la subhoja «income gap». En la resolución de este ejercicio se mostrará además la diferencia al establecer una línea de pobreza absoluta o relativa. Comencemos estableciendo una línea relativa, como porcentaje de la media. Como los distintos escenarios hacen variar la distribución de renta, también lo hace la media total, y por tanto el valor de la línea. Si se utiliza la distribución de rentas de partida, el valor de *I* es el que aparece reseñado en las celdas C110 ó C111, según se utilice la expresión primera o segunda mostrada en [3]. El p*overty gap, (PG)* se obtuvo en el ejercicio anterior, por lo que no se vuelve a explicar su cálculo que se muestra en D106. Solamente es necesario calcular el número de pobres, (que resulta de sumar los indicadores iguales a 1 cuando la unidad es pobre, que aparecen en la columna C y cuya suma total se sitúa en C106) y la renta media entre los pobres. Para obtener la renta media entre los pobres, se añade una columna más, a la derecha del déficit, que contiene el producto de la renta por el indicador 1 ó 0 que indica si se es o no pobre. Con ello se consigue sumar exclusivamente las rentas de los pobres en la columna E, para obtener la media en E103. El valor de *I* se ofrece siempre de dos formas alternativas, lo que puede servir de comprobación de que el ejercicio está bien realizado. En la casilla C110 (=D106/(B108*C103) se obtiene *I* como el cociente entre el *PG* y el producto de la línea de pobreza por el n.º de pobres). En la casilla C111 (=1-(E103/B108)) se calcula como la unidad menos el cociente entre la renta media de los pobres y la línea de pobreza. Los cálculos son idénticos para el resto de escenarios, por lo que explicamos los cambios en la distribución para después comentar los resultados.

Para obtener la distribución del escenario *a)*, simplemente se hace nula la renta de las 26 primeras unidades, que son las pobres en la distribución inicial. Ello conlleva un descenso en la renta media, por lo que al establecer una línea relativa el valor de ésta baja. En el escenario *b)* se duplican todas las rentas, por lo que la renta media se eleva, y también la de los pobres. Los valores de la distribución de rentas modificados respecto de la situación inicial se marcan en azul para que se distingan con claridad. Si el lector desea cambiar los valores de la línea de pobreza debe se consecuente con los valores de la distribución y comprobar que el número de pobres se recalcula correctamente.

Los resultados obtenidos se muestran en el siguiente cuadro de la misma forma que aparecen en la hoja de resultados:

CUADRO 3.2.3.1. *Resultados del ejercicio 3. Índices de pobreza en tres escenarios alternativos*

	Inicial	*Escenario a)*	*Escenario b)*
Línea relativa (70% media)	**3828.89**	**3550.06**	**7657.77**
N.º de pobres	26	26	26
Renta media de pobres	1532.01781	0	3064.03561
PG	59718.56	92301.52	119437.12
I	0.59987893	1	0.59987893
Línea absoluta	**3828.89**	**3828.89**	**3828.89**
N.º de pobres	26	26	16
Renta media de pobres	1532.01781	0	1478.5039
PG	59718.5622	99551.03	37606.11
I	0.59987893	1	0.61385529

Los resultados anteriores se presentan primero para líneas de pobreza relativas del 70% de la media, lo que supone un valor diferente en cada escenario. Con respecto a la situación inicial, el que todos los pobres pasen a tener renta nula implica un *PG* mayor (92301>59718). Además, si todos los pobres cuentan con renta nula, el valor de I es el máximo que se puede alcanzar, la unidad, ya que su carencia total de renta (o *PG*) coincide con *qz*, y por tanto *PG* entre *qz* es la unidad. En ele scenario en que se duplican las rentas, el número de pobres sigue siendo 26, pero la renta media entre los pobres se duplica, y también el *PG*, por lo que el valor de I permanece inalterado con respecto a la situación inicial (0.59988=0.59988).

Cuando se considera la misma línea de pobreza para todos los escenarios, el valor de referencia es el inicial, de 3828.89 unidades monetarias, y ello implica modificaciones en el número de pobres del escenario b), que pasan a ser 16, ya que las rentas se duplican y la línea de pobreza baja respecto a la situación en que se fijaba una línea relativa. Cuando todos los pobres cuentan con renta nula, el valor de *I* sigue siendo el másximo posible, 1, a pesar de que haya habido modificaciones en la distribución, como muestra el hecho de que el *PG* crece con respecto a la situación en que se fijaba la línea de forma relativa (ahora la línea es mayor 3828.89>3550.06) o con respecto a la situación de partida, ya que los pobres (véase *PG*) son más pobres (59718.56<99551.03).

Es interesante notar que el máximo valor de *I* igual a la unidad se puede presentar en escenarios muy variados con distinto número de pobres, de líneas de pobreza y valores de *PG*, pero siempre se estará reflejando el hecho de que para el contexto que sea, «ya no se puede ser más pobre de lo que son los pobres», debido a que el escenario será tal que todos los pobres no tendrán nada. De igual forma, se puede interpretar el hecho de que no haya nadie por debajo de la línea de pobreza como la situación más favorable, y en tal caso, *I* valdrá siempre 0, porque el *PG* será nulo porque no hay nadie con menos renta que la marcada por la línea de pobreza. Esto puede ocurrir en muchos escenarios alternativos, porque como ya habrá quedado claro, todo depende de dónde se fije el valor de la línea *z*.

7.3.2.4. Income gap ratio

En aras de compensar la insensibilidad a la intensidad de la pobreza o a la proporción de pobres que presentan los índices anteriores, se suele calcular el *poverty gap ratio (HI)* que se obtiene como el producto de *H* por *I*, y representa el cociente entre la brecha de pobreza y *Nz*.

$$HI = \frac{\sum_{i=1}^{q} z - x_i}{Nz} \qquad [4]$$

La interpretación de *HI* es similar a la de *I,* representa la proporción entre la renta que habría que transferir a los pobres para que dejasen de serlo y la renta que tendría *toda la población* si todo el mundo se situase sobre la línea de pobreza.

Es decir, ambos índices difieren en que *H* calcula la proporción de renta para que los pobres dejen de serlo sobre la renta si *todos los pobres* se situasen sobre la línea de pobreza, mientras que *HI* calcula la proporción de renta para que los pobres dejen de serlo sobre la renta total si *toda la población* estuviese sobre la línea de pobreza. Siempre que ocurra que la población total no es pobre, habrá un número mayor de población total que de pobres ($N>q$ y $q/N<1$) y por ello en genral $H<HI$. De hecho $HI=Hq/N$ y por eso siempre $HI<=H$. Comparar ambas medidas tiene sentido solamente si hace con coherencia: la pobreza es la misma se mida a través de *H* o a través de *HI*, pero el peso de la transferencia para acbar con la pobreza medida sobre la renta de toda la población o solamente de los pobres si se situasen sobre la línea de pobreza, no puede pesar lo mismo.

EJERCICIO 4: Calcular *HI* con la distribución de renta existente, y comprobar su valor en los siguientes escenarios alternativos:

a) Todos los pobres tienen renta nula.

b) No hay nadie pobre (todos los pobres se sitúan sobre la línea de pobreza).

c) Toda la población es pobre (la línea de pobreza supera la renta máxima).

d) Toda la población tiene renta nula.

Todos los cálculos previos necesarios para obtener *HI* se han desarrollado en los ejercicios anteriores (cálculo de *H* y de *I*) por lo que lo único que se ha de explicar en la subhoja «poverty gap ratio» son los cambios relativos a las distribuciones (que se colorean en azul) para comparar los resultados obtenidos en los distintos escenarios. El cuadro siguiente resume todos los valores obtenidos (para una línea de pobreza del 70% de la media de la distribución inicial, salvo en *c)* donde para hacer a todos pobres se sitúa *z* por encima de la renta máxima):

CUADRO 3.2.4.1. *Resultados del ejercicio 4. H, I y HI en tres escenarios alternativos*

	Inicial	*a)*	*b)*	*c)*	*d)*
H	0.26	0.26	0	1	1
I	0.6	1	0	0.45	1
HI	0.16	0.26	0	0.45	1

A partir de la situación inicial en la que *HI* toma valor 0.16, si los pobres son los mismos pero su pobreza es más intensa (como indica el hecho de que su renta es nula en el escenario *a*), el valor de *HI* se eleva hasta 0.26. Aunque los pobres son muy pobres, en particular lo máximo que lo pueden ser, no toda la población es pobre, solamente el 26%, por lo que no es esta la peor situación imaginable. El mejor escenario que capta *HI* se refleja en *b)* y es aquél en que no hay nadie pobre (independientemente de la renta que tengan, con tal de que se sitúen por encima de *z*), por lo que tanto *H* como *I* son nulos y también su producto. Otra situación extrema es captada en el escenario *c)* donde toda la población es pobre, pero como no es todo lo pobre que se puede ser (contar con renta nula) *HI* no toma el valor máximo, 1, reservado para el escenario *d)* Todo el mundo es pobre y lo más que se puede ser: contar con renta nula.

7.3.2.5. Índices de Foster Greer y Thorbecke

Hasta el momento, ninguno de los índices considerados incorpora la medición de la desigualdad existente entre la población pobre, lo cual no quiere decir que no existan tales índices [3]. Mostramos a continuación el que creemos que es el índice más utilizado en el trabajo empírico, FGT propuesto por Foster Greer y Thorbecke (1984) y que supone una generalización de los índices anteriores. Se calcula como:

$$FGT(\alpha) = \frac{1}{N}\sum_{i=1}^{q}\left[\frac{z - x_i}{z}\right]^{\alpha} \qquad [5]$$

El parámetro a incorpora preocupación o aversión a la pobreza, de modo que valores mayores del mismo implican que se otorga un mayor peso a los más pobres [4]. Esto es así porque para niveles mayores del parámetro se está otorgando un peso superior a los déficits mayores de renta, que son los que aparecen en las unidades más pobres, o que más distan de la línea de pobreza. Otra forma de interpretar es valor del parámetro de aversión a la pobreza es considerar la importancia de las transferencias hacia los individuos pobres: cuanto mayor es el valor de a, mayor importancia reconoce el índice a transferencias progresivas hacia los más pobres.

Se puede pensar cómo contribuye cada uno de los pobres a la medición de la pobreza en función del valor del parámetro a: por ejemplo, si toma valor nulo, el valor del índice FGT coincide con *H*. Es decir, calculamos el porcentaje de pobres, pero sin tener en cuenta la intensidad de la pobreza ya que lo único que se hace es contar el número de pobres. La cuantía de los déficits de cada uno de los pobres no es tenida en consideración, y cada pobre cuenta como una unidad en el total, que se divide entre la población total. Si por el contrario la aversión a la pobreza vale uno, se calcula el *poverty gap ratio HI*, que sí es sensible a la intensidad de la pobreza, ya que en este caso, cada pobre contribuye al índice dependiendo de cuánto vale su déficit hasta alcanzar la línea de pobreza, de manera que se considera con mayor importancia a los pobres cuanto más pobres son. Para a igual a infinito (lo que en la práctica puede significar una cifra no superior a 100), se estaría considerando exclusivamente a la unidad más pobre entre las pobres, lo que implica la máxima aversión a la pobreza.

---

[3] Pueden citarse el índice de Sen o el de Thon como ejemplo más comunes.

[4] Es común interpretar el parámetro de aversión a la pobreza de forma que cuanto mayor es éste, mayor es el peso que se otorga al bienestar de la unidad más pobre. Pero también es cierto que a medida que se consideran valores superiores del parámetro se otorga más peso a cualquier pobre en comparación con otra unidad menos pobre, lo que puede dar lugar a que las ordenaciones al comparar distribuciones alternativas sean ambiguas.

EJERCICIO 5: Calcular FGT(0), FGT(1), FGT(2) para la distribución existente y para una en la que todos tuvieran el doble de la renta (manténgase el valor absoluto de la línea de pobreza y compruébese la variación de FGT a medida que se modifica el valor de la aversión a la pobreza).

SOLUCIÓN: Se presenta en la subhoja FGT de la hoja de ejercicios. Para calcular los índices FGT es necesario añadir una columna en la que se calcule para cada observación de la distribución de renta el cociente entre el déficit respecto de la línea de pobreza elevado al parámetro de aversión que se desee. Ello se obtiene en la columna E «gaps normalizados y elevados» para la distribución inicial. El parámetro de aversión se fija en la casilla B110 y automáticamente se producen los cambios convenientes en la columna E (=SI(C3=1;(D3/B$108)^B$110;0)). Lo que implica esta expresión es que si la unidad considerada es pobre (C3=1) se calcula el ratio entre su déficit (D3) y la línea de pobreza (B108) elevada al parámetro de aversión a la pobreza (B110). En caso contrario, el gap elevado es nulo. La suma de los gaps elevados para todas las unidades pobres, dividida entre el total de observaciones (E103) entre el total de población (A102) es lo que se obtiene en E112, el valor del índice FGT buscado. Para la distribución inicial, las columnas utilizadas son de la A a la E, mientras que para la distribución de rentas duplicadas, se replica el proceso entre las columnas H a L, con el valor del índice FGT en la casilla L112. Como obtener los índices FGT para distintos parámetros de aversión a la pobreza es relativamente sencillo (solamente requiere modificar el parámetro y apuntar el nuevo valor obtenido), se presentan en el siguiente cuadro los resultados para los valores de a que propone el ejercicio y para algunos más:

CUADRO 3.2.5.1. *Valores de FGT para diferente aversión a la pobreza*

Valor de α	*Distribución inicial*	*Distribución duplicada*
0	0.19	0.12
0.25	0.167	0.107
0.5	0.149	0.097
0.75	0.135	0.089
1	0.124	0.082
1.25	0.114	0.077
1.5	0.107	0.077
2	0.095	0.065
2.5	0.087	0.059
3	0.080	0.055

Nota: Todos los cálculos se realizan para una línea de pobreza de z=2735 (50% de la media de la distribución inicial)

De los resultados anteriores cabe destacar lo siguiente:

a) Cuando la aversión a la pobreza es nula, FGT(0) coincide con la proporción de pobres sobre el total, *H*, que son 19 sobre un total de 100 en la distribución inicial, y 12 sobre el mismo total al duplicar las rentas, (lo que explica que el valor caiga).

b) Si a toma valor unitario, se obtiene el poverty gap ratio (*HI*). El lector puede comprobar que los resultados se igualan sin más que modificar la línea de pobreza para que sean iguales en las subhojas de headcount ratio, poverty gap ratio y FGT.

c) Para cualquiera de los valores de aversión a la pobreza se produce una ordenación no ambigua en el sentido de que siempre se capta una pobreza mayor en la distribución inicial que en la duplicada, lo que se explica porque los pobres son menos pobres, y la línea de pobreza se mantiene en términos absolutos (2735 unidades), no se duplica.

d) Si la línea de pobreza se duplicase, el valor de la pobreza sería el mismo para cualquiera de los FGT calculados, lo que revela que estos índices cumplen invarianza ante cambios de escala.

e) Cuanto mayor es el parámetro de aversión a la pobreza, menor es el valor numérico del índice. Ello no quiere decir que la pobreza descienda. La pobreza es siempre la misma, pero se mide de forma diferente. Lo verdaderamente relevante es qué ocurre cuando se altera el valor de la aversión a la pobreza al comparar dos distribuciones diferentes, y como se comentó en c), la ordenación siempre indica mayor pobreza para la distribución inicial que para la duplicada.

7.4. LAS CURVAS *TIP*

Además de los índices expuestos previamente, una herramienta de uso generalizado en el análisis empírico de la pobreza son las curvas TIP (Three I s of Poverty: Incidence, Intensity and Inequality, Jenkins and Lambert (1997)). Se trata de curvas construidas a partir de la agregación de déficits de renta (o de la dimensión —única— que se considere) para los acumulados de población ordenados por nivel de pobreza. Son similares a las curvas de Lorenz en al análisis distributivo, y cuentan como ventaja fundamental la incorporación de las consideraciones de incidencia, intensidad y desigualdad. Además, las ordenaciones que se derivan de comparar las curvas TIP de dos distribuciones diferentes son consistentes con las que se derivarían de una amplia clase de índices y umbrales de pobreza. Las curvas TIP se calculan como sigue:

1) Se ordena a las unidades pobres (es decir bajo la línea de pobreza *z*) de más pobre a menos pobre, y se obtiene el déficit con respecto a la línea de cada unidad. Si la unidad no es pobre, su déficit sería positivo, por lo que no lo consideramos. Se construye entonces un vector de déficits D_x sobre la distribución digamos de renta, *x* como:

$$D_x = \max\{0,(z-x_i)\}$$

2) Se agregan los déficits para cada una de las unidades consideradas pobres, de manera que si incluimos 2 pobres sumamos 2 déficits. Si incluimos 10 pobres, 10 déficits, y si incluimos al total de pobres, tendríamos lo que denominamos *PG* en la ecuación [2].

$$Acumulado\,D_x = \sum_{i=1}^{q} D_i$$

En la expresión anterior se van sumando los déficits hasta la observación *q*, que sería la última por debajo del umbral de pobreza *z*. La unidad siguiente no sería pobre y su déficit sería nulo.

3) La expresión matemática de una curva TIP es la siguiente:

$$TIP(D_x, p) = \frac{\sum_{i=1}^{q} D_i}{N}$$

Nótese que *q* es el número total de pobres, *N*, el total de población, y *p* hace referencia al 100·*p* por ciento de los individuos más pobres, con $0 \leq p \leq 1$. Para cada valor de *p*, $TIP(D_x; p)$ representa el gap acumulado por el 100·*p* por ciento más pobre de la población, dividido entre el total de individuos *N*.

También pueden construirse curvas TIP a partir de un vector de déficits normalizados, al dividir cada uno de los déficits por el valor de la línea de pobreza *z*. En tal caso, cada déficit normalizado se calcula como:

$$\varphi x_i = \frac{D_{x_i}}{z} = \frac{z-x_i}{z}$$

y la expresión de la curva TIP sería ahora:

$$TIP(\varphi_{x_i}, p) = \frac{\sum_{i=1}^{q} \varphi_{x_i}}{N}$$

En una curva TIP, la incidencia de la pobreza queda reflejada en el punto en que se hace horizontal, es decir, cuando no agregamos más unidades porque superan el umbral de pobreza. En tal punto, conocemos la proporción de población total que es pobre. La intensidad del fenómeno se refleja en la altura máxima alcanzada, ya que la construcción se realiza a partir de la agregación de todos los déficits de los pobres. (Se incorpora además la medición de la desigualdad entre los pobres mediante la concavidad de la curva TIP). La representación gráfica puede aclarar algo estas ideas:

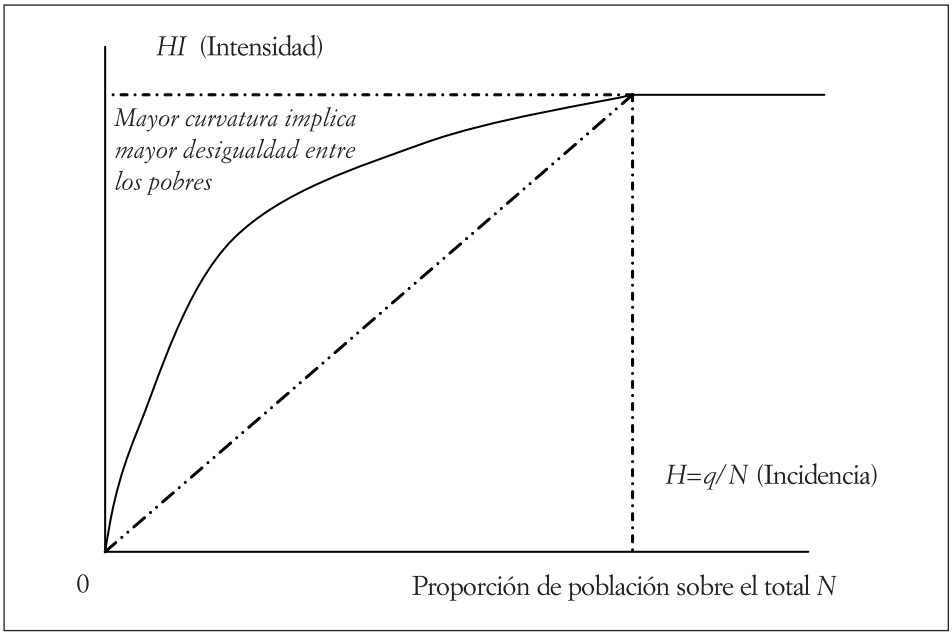


GRÁFICO 4.1. *Curva TIP normalizada: incidencia, intensidad y desigualdad de la pobreza*

En el eje de ordenadas se representa la proporción acumulada de población. En el límite, si toda la población fuese pobre, la curvatura antes de llegar al tramo horizontal se extendería hasta la unidad, si bien es que lo normal es que *H* tome un valor inferior a la unidad. La altura máxima de la curva TIP coincide (si se agregan déficits normalizados) con el índice *HI*, que captura la intensidad del fenómeno. La curva no puede elevarse más allá, puesto que sucesivas unidades no serían pobres, por tanto sus déficits son nulos. En cuanto a la curvatura, cuanto más tiende la curva a una L invertida, mayor es la desigualdad entre los pobres. En un extremo, si todos los pobres tuviesen renta nula excepto una sola unidad, la menos pobre que tuviera una cantidad cualquiera positiva e inferior a *z*, la curva TIP sería vertical hasta la altura *HI*, y después horizontal. En otro extre-

mo si tuviéramos una línea recta, indicaría que todas las unidades por debajo del umbral de pobreza cuentan con la misma renta.

EJERCICIO 6. Calcule la curva TIP normalizada de la renta, y compare su resultado con una distribución en que todos los pobres contaran con la renta media de entre los pobres.

SOLUCIÓN: Los resultados de este ejercicio se muestran en la subhoja curvas TIP. En primer lugar construimos la curva TIP para la distribución inicial. A partir de la columna E, donde se muestran los déficits con respecto a la línea de pobreza (establecida en B108), se calculan en la columna F los gaps normalizados (=SI(C3=1;(E3/B$108);0)). Es decir, si la unidad es pobre, se calcula el cociente entre su déficit y la línea de pobreza, y si no es pobre, se supone valor nulo. En la columna G se agregan los gaps normalizados unidad a unidad, y en la H, se divide este acumulado entre la población total, para obtener la variable que se representa en ordenadas (se repite en la columna L para marcar posteriormente el área que se grafica). Para calcualr el acumulado de población (eje de abscisas de la curva TIP) se calcula en la columna K el cociente entre las observaciones acumuladas y el total de observaciones. Una vez calculados estos valores de las columnas K y L, se marca el rango de datos para escoge la opción «dispersión» que permite obtener el siguiente gráfico:

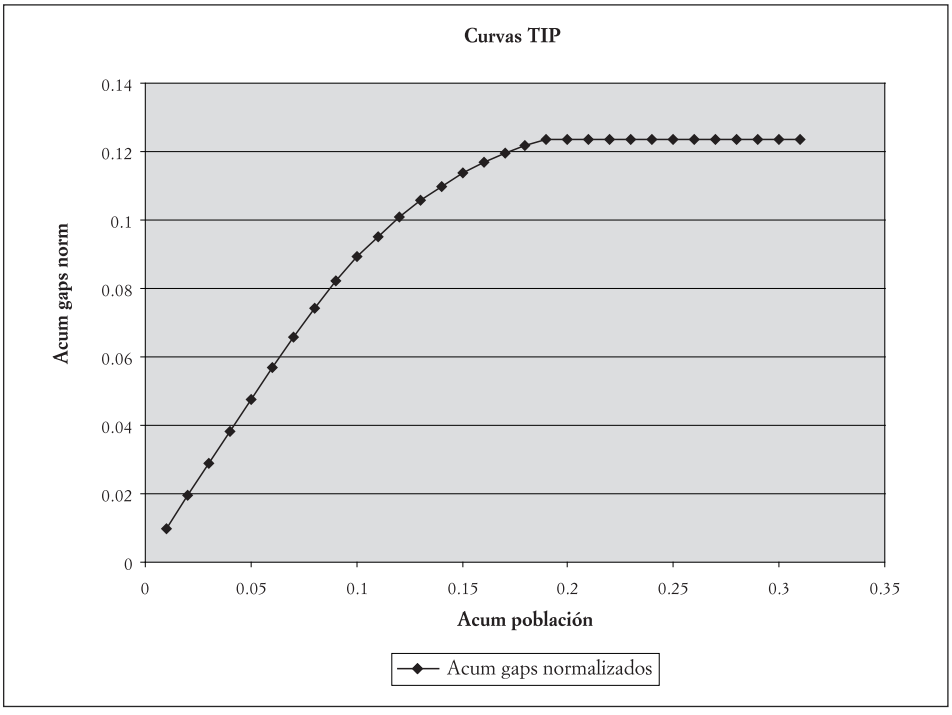


GRÁFICO 4.2. *Solución al ejercicio 6. Obtención de la curva TIP*

Como se puede apreciar (mejor si se acude a los datos originales) el acumulado de población para el cual la curva alcanza su tramo plano es 0.19, exactamente el tanto por uno de pobres (*H*). El valor corresponsiente para el acumulado de gaps normalizados, o altura máxima de la curva TIP se alcanza en el valor 0.1237, lo que coincide con el valor de HI. Esos dos valores captan así la incidencia de la pobreza y la intensidad de la misma. El hecho de que la curva TIP sea estrictamente cóncava indica la desigualdad entre los pobres.

Precisamente, la segunda parte del ejercicio trata de ilustrar el efecto de la desigualdad entre los pobres sobre la concavidad de la curva TIP. Al proponer una distribución en la que todos los pobres cuentan con la renta media de los pobres de la distribución inicial, no varían ni la incidencia (*H* sigue siendo 0.19) ni la intensidad del fenómeno (*HI*=0.1237), puesto que el número de pobres es el mismo y el povery gap es también idéntico, aunque los gaps individuales se distribuyan ahora de forma uniforme. El nuevo gráfico, que se construye a partir de la repetición de los cálculos anteriores en las columnas W hasta AH, es el que se muestra a continuación:

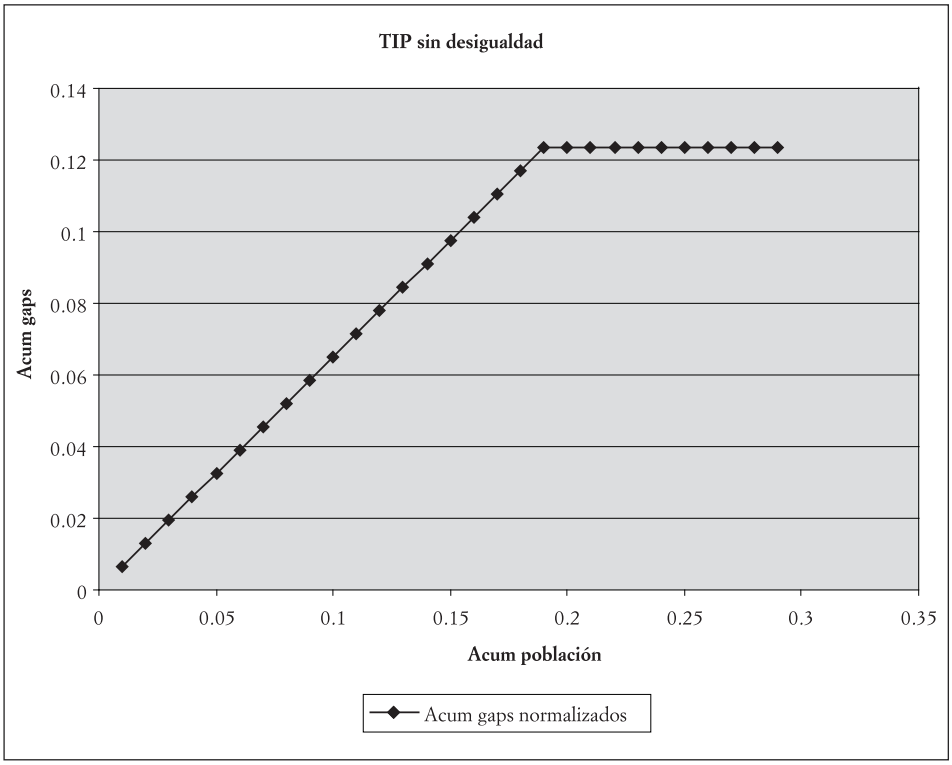


GRÁFICO 4.3. *Solución al ejercicio 6. Obtención de la curva TIP sin desigualdad entre los pobres*

Como se puede observar, el valor en ordenadas y abscisas del punto en que no se agregan más gaps sigue siendo (X,Y)=(*H, HI*)=(0.19,0.1237) si bien la curva TIP es ahora «cuasicóncava» y en particular, no presenta curvatura ninguan debido a que no existe desigualdad entre los pobres.

Una última cuestión de interés en relación con las curvas TIP es la dominancia. Si una curva TIP se sitúa para todo el dominio en el que la curva no alcanza su tramo horizontal por encima de otra, diremos que existe dominancia TIP (dominancia estocástica de primer orden), y por tanto, la pobreza es mayor para cualquier línea de pobreza común considerada por debajo de la existente. El hecho de que algunos índices puedan no reflejar la dominancia TIP no implica un fallo de las curvas TIP, sino que algunos índices no cumplen las características que sería deseable que cumpliesen. Piénsese qué ocurriría si a partir de una distribución A aumentara la desigualdad entre los pobres construyéndose una distribución B cuya curva TIP sería dominada por la de A: si el número de pobres no ha variado, un índice como *H* no lo reflejaría.

Cuando se producen cruces entre las curvas TIP, ya no es tan fácil establecer conclusiones válidas en cualquier circunstancia, pero sí se pueden establecer algunas conclusiones si se fija la aversión a la pobreza. Por ejemplo, cuando las curvas TIP se cruzan una sola vez, y la desigualdad medida a través de FGT(2) es superior en la curva TIP que empieza por encima, se puede decir que en esa distribución hay más pobreza cuando se asume una aversión a la desigualdad de la renta de los pobres suficientemente elevada.

Las comparaciones entre curvas TIP son particularmente interesantes para el mismo entorno en momentos diferentes del tiempo, para distintos territorios pertenecientes a un conjunto mayor, o para hogares de composiciones diferentes. Según Jenkins y Lambert (1997), cuando hay dominancia clara TIP de una distribución sobre otra, existe además margen para bajar el valor de la línea de pobreza de la distribución más pobre y mantener la dominancia TIP. Si se baja el umbral de pobreza, en principio hay menos pobres, y la suma de los gaps de pobreza también es menor. Si incluso así la distribución dominada en sentido TIP sigue siendo más pobre es que la comparación en términos de pobreza no presenta dudas. El margen para mover la línea de pobreza será mayor cuanto mayor sea la distancia vertical entre las dos curvas TIP comparadas y cuanto más cercana esté la TIP que denota mayor pobreza a la L invertida que significaría pobreza máxima.

7.5. LA POBREZA COMO FENÓMENO MULTIDIMENSIONAL

La pobreza es un fenómeno multidimensional, pero tiene interés el reducir su medida a una sola dimensión a través de un índice. Esto no es tarea fácil, y la literatura relativa a esta cuestión es todavía muy reciente. Véa-

se Tsui (2002) y Bourguignon y Chakrawarty (2002). Conscientes de la multidimensionalidad del fenómeno, se pueden calcular distintos índices de pobreza sobre variables diferentes, pero ello puede llevar a situaciones no comparables. Esto ocurre porque cuando se calcula un índice unidimensional basado por ejemplo en la renta, la ordenación que se deriva de su cálculo es completa, o dicho de otra forma, se pueden comparar todas las distribuciones de renta y establecer una ordenación de más a menos pobreza. Pero cuando se utiliza una familia de índices, la ordenación que resulta puede ser parcial o incompleta, apareciendo situaciones no comparables.

Nos encontramos entonces en una situación en que la unidimensionalidad permite las comparaciones, pero no capta toda la amplitud del fenómeno que está midiendo. Siguiendo a Amartya Sen (1987) el bienestar procede de las capacidades para funcionar en la sociedad, y la pobreza vendría dada por la ausencia de tales capacidades. La pobreza surge por tanto en circunstancias de insuficiencia de renta o educación, mala salud, ausencia de libertad o de derechos fundamentales, y ello revela que se trata de un fenómeno multidimensional y que se mide de forma incompleta cuando se considera exclusivamente la carencia de renta.

7.6. CONSIDERACIONES FINALES

Hemos dejado constancia de que la medición de la pobreza es importante para poder luchar contra la misma, pero no es fácil encontrar referencias bibliográficas sencillas y que aborden su estudio desde el nivel más básico, y por ello se ha planteado la elaboración de este capítulo. Una de las principales razones por las que la medición de la pobreza no está tan desarrollada como otras cuestiones (como por ejemplo, la desigualdad) dentro de la Economía del Bienestar, es que quienes son pobres están enmarcados en los colectivos olvidados y no constituyen grupos de presión que puedan luchar por sus propios intereses. Ello no quiere decir que la pobreza sea un problema cuantitativa o cualitativamente poco importante ni de resolución sencilla. La propia definición de pobreza puede darse en unas pocas palabras, pero su materialización en cifras no es fácil, pues traducir a un solo número lo que está condicionado por múltiples causas no es inmediato. Así, aunque la pobreza constituye un fenómeno multidimensional, su medición se ha limitado a una sola dimensión, normalmente la renta. La elección de una sola dimensión tampoco está exenta de problemas, pues no está claro que la renta esté bien definida o que los datos permitan medirla con exactitud. Incluso aceptando un enfoque unidimensional, no es válido utilizar cualquier índice, pues hay propiedades de la medición a las que no conviene renunciar. Una buena opción que es al mismo tiempo sencilla es utilizar los índices de FGT para niveles de aversión a la pobreza superiores a 2, lo que permite estar seguros que se está usando una medida consistente. La cuan-

tificación de la pobreza puede acompañarse de la representación de curvas TIP, que permiten incorporar las dimensiones de incidencia, intensidad y desigualdad de la pobreza. Las tendencias más recientes de investigación en pobreza avanzan en la línea de proponer índices multidimensionales, si bien queda mucho camino por recorrer.

Para terminar, presentaremos algunas direcciones de sitios web que puedan resultar de interés para el lector que pretenda ahondar más en el estudio de la pobreza de manera autodidacta:

— El ISER (Institute for Social and Economic Research) cuenta con una página web en la que se presentan distintos eventos que pueden ser de interés en el estudio de la pobreza y materias relacionadas. También cuenta con una colección de documentos de trabajo y otras publicaciones entre las que existen muchos trabajos relacionados con medición de la pobreza: http://iserwww.essex.ac.uk/search/catsearch/publications.php. También vinculado al ISER puede accederse a EUROMOD, modelo de microsimulación tax benefit aplicado a una gran cantidad de países europeos http://www.iser.essex.ac.uk/msu/emod/

— El Banco Mundial cuenta entre un amplio abanico de materias de estudio, un apartado de «Análisis de la pobreza» en el que se puede encontrar mucha información de interés, desde noticias, material didáctico, resultados de la comparativa de medición de pobreza en diferentes países, documentos de trabajo, y un largo etc. http://wbln0018.worldbank.org/LAC/LAC.nsf/ECADocByUnid2ndLanguage/974C7C710AE6C4AE85256C60005ED4BD?Opendocument. Existe además dentro del Banco Mundial un sitio Web dedicado enteramente a pobreza, PovertyNet, en el que se ofrece material muy interesante acerca de cuestiones como: análisis de la pobreza, pobreza y salud, pobreza y desigualdad, pobreza y economía sumergida, estrategias de reducción de la pobreza, empobrecimiento, y otras materias relacionadas. http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/0,,menuPK:336998~pagePK:149018~piPK:149093~theSitePK:336992,00.html

— La Universidad de Laval pone a disposición del público en general un microsimulador DAD (Software para análisis distributivo) diseñado para realizar análisis comparativo de bienestar, desigualdad, pobreza entre distribuciones. Está disponible en: http://132.203.59.36/DAD//index.html o bien en: http://132.203.59.36/DAD/index2.htm

— El The International Poverty Centre (IPC) es un proyecto conjunto del programa de desarrollo de las Naciones Unidas y el gobier-

no de Brasil especializado en el análisis de la pobreza y la desigualdad. Su página web contiene información relativa a seminarios y publicaciones de interés: http://www.undp-povertycentre.org/

— El Instituto de Estudios Fiscales Británico ofrece una página web con publicaciones relativas a desigualdad, pobreza y bienestar en la dirección: http://www.ifs.org.uk/publications.php?heading_id=8

— Otras páginas web en las que se puede encontrar información interesante relativa a informes, seminarios y publicaciones son (por citar sólo algunas) la del Banco Interamericano del desarrollo (http://www.iadb.org/). Human Development Reports (http://hdr.undp.org/). La Comisión Económica para América Latina y el Caribe, CEPAL (http://www.eclac.cl/) y desde este sitio web, sus enlaces proporcionan mucho más vínculos interesantes (http://www.eclac.cl/Enlaces/)

BIBLIOGRAFÍA

Bourguignon, F. and Chakravarty, S.R. (2002) «Multidimensional poverty orderings», *Delta Working Papers 2002-22, Delta*.

Foster, J., J. Greer, and E. Thorbecke (1984). «A Class of Decomposable Poverty Measures», *Econometrica*, 42 (3), pp761-766.

Gradín, C. y Del Río, C. (2001), *Desigualdad, pobreza y polarización en la distribución de la renta en Galicia*, Instituto de Estudios Económicos de Galicia - Fundación Pedro Barrié de la Maza (ed.), vol. 11, A Coruña.

Jenkins, S.P. and P. J. Lambert (1997) «Three 'I's of Poverty Curves, with an Analysis of UK Poverty Trends», *Oxford Economic Papers*, New Series, Vol. 49, No. 3, pp. 317-327.

Ravallion, Martin, G. Datt, and D. van de Walle (1991). «Quantifying Absolute Poverty in the Developing World», *Review of Income and Wealth*, vol. 37(4), pages 345-61, December.

Sen, A (1976), «Poverty: An Ordinal Approach to Measurement», *Econometrica*, Vol. 44, N.º 2, pp. 219-231.

Sen, A. (1987) *On Ethics and Economics.* Oxford: Basil Balckwell.

Tsvi, K. Y. (2002) «Multidimensional poverty indices». *Social Choice and Welfore.* Springer, vol 19 (1) pages 69-93.

CAPÍTULO VIII
EVALUACIONES DE BIENESTAR

JORGE ONRUBIA FERNÁNDEZ

8.1. INTRODUCCIÓN

En el análisis microeconómico, una forma empleada frecuentemente para decidir hasta qué punto es recomendable una determinada política pública es medir su impacto sobre el bienestar tanto individual como social. En el caso del bienestar individual, se trata de obtener una valoración monetaria del cambio de bienestar (mejora o empeoramiento) que obtendría el beneficiario directo de esa actuación. En el segundo caso, la evaluación de bienestar social tiene como misión establecer una ordenación de preferencias para la sociedad respecto de políticas alternativas, mediante una valoración simultánea de su impacto en el nivel de renta y en su distribución.

El concepto de bienestar individual atiende exclusivamente a la situación del ciudadano u hogar que se ve afectado por esa actuación del sector público. Por tanto, su valoración se identifica generalmente con una cuantificación de la utilidad que le reporta el consumo del correspondiente bien o servicio público o el importe de la prestación monetaria recibida. La comparación entre mecanismos de intervención alternativos exige plantear esta evaluación en términos relativos, situándonos ante la medición del cambio en el bienestar individual provocado por la sustitución de una política por otra. Generalmente, la medición del bienestar individual, en la medida que afecta en exclusiva a la utilidad de un único agente, es una forma de cuantificar el criterio de eficiencia de Pareto en lo referente a sus decisiones de consumo, ahorro u oferta de trabajo.

El concepto de bienestar social, por su parte, está íntimamente ligado a la evaluación de los resultados distributivos que genera el sistema de mercado y/o las diversas intervenciones del sector público. Por consiguiente, su cuantificación obliga a considerar a la totalidad de individuos de la sociedad (o población de referencia, en cada caso). Su especificación a través de una forma funcional matemática, generalmente conocida como función de bienestar social, suele identificarse con la función objetivo del decisor social, entendiendo como tal al legislador, gobierno o cualquier administración pública que adopta decisiones en interés colectivo.

Al incorporar esta perspectiva «social», el criterio de eficiencia paretiano deja fuera las consideraciones acerca de la valoración qué merece a la sociedad la forma en la que se distribuye el bienestar entre los ciudadanos. De hecho, su aplicación como criterio de evaluación social recomendaría como deseable cualquier conjunto de asignaciones económicas en las que el valor global de las mismas fuese superior al de otra combinación alternativa, incluso en aquellos casos en los que se produjese un aumento en la desigualdad con la que se distribuyen los recursos económicos determinantes del bienestar individual. Esto sucedería, por ejemplo, si el bienestar del hogar más rico de la sociedad aumentase en 100 unidades, permaneciendo inalterado el bienestar del resto de hogares. Incluso, también resultaría preferible si esa mejora en el bienestar del hogar más rico se produjese simultáneamente con una reducción de 20 unidades en el bienestar del hogar más pobre. Sencillamente, el criterio de eficiencia *paretiana* se limita a sumar las «porciones del pastel» obtenidas por los distintos individuos como resultado de las asignaciones generadas por el mercado con o sin la actuación de los gobiernos, sin entrar a valorar si existe una mayor o menor desigualdad entre estas porciones.

Por consiguiente, la cuestión central que plantea la evaluación del bienestar social es cómo agregar el bienestar de los distintos individuos (u hogares) cuando aceptamos que la sociedad presenta *aversión hacia la desigualdad*. Como veremos en este capítulo, la manera en la que se suman las utilidades de los individuos para construir las preferencias de la sociedad resulta fundamental para conseguir una valoración conjunta de la eficiencia de las asignaciones y de la distribución de sus resultados.

La medición del bienestar social utiliza generalmente la renta de los individuos (o de los hogares) como variable fundamental, la cual se ve modificada, en cada caso, en los importes correspondientes a las prestaciones recibidas, los impuestos pagados, o la valoración de los bienes o servicios consumidos. En el análisis empírico, el enfoque individualista es el más empleado, y exige disponer de información a nivel de microdatos para cada unidad de análisis, si bien en ocasiones, también se utilizan especificaciones abreviadas del bienestar. Esta aproximación es consistente con la propiedad de individualismo que caracteriza a aquellas funciones de bienestar social en las que la evaluación del decisor social respeta las valoraciones particulares que hacen los individuos de su propio bienestar.

En este capítulo presentamos en primer lugar los fundamentos metodológicos empleados habitualmente en la medición del bienestar individual, con algunos ejemplos que ilustran su aplicación en el terreno empírico. En segundo lugar, se exponen los principales métodos para llevar a cabo evaluaciones de bienestar social, con especial atención a los criterios de dominancia estocástica, acompañados también de ejemplos de aplicación.

8.2. LA MEDICIÓN DEL BIENESTAR INDIVIDUAL

Para evaluar la situación en la que se encuentra un individuo en relación con los recursos a los que tiene acceso, identificamos la utilidad que le reporta bien su renta disponible bien los consumos realizados con su nivel de bienestar individual. Si tenemos en cuenta que cualquier reasignación de estos recursos, ya sea consecuencia del funcionamiento del mercado ya esté originada por la intervención del gobierno, afectará a este bienestar individual, conocer en qué medida un individuo afectado por este cambio mejora o empeora su situación requiere cuantificar la utilidad alcanzada antes y después del cambio asignativo. El método más habitual consiste en establecer una medida monetaria del bienestar individual, de forma que puedan realizarse comparaciones homogéneas entre situaciones alternativas, en términos de valor económico. Además, este cómputo monetario de los cambios de bienestar permite asociar su medida con la predisposición al pago que mostraría un individuo por acceder a la mejora de bienestar que le reportaría la política adoptada.

El excedente del consumidor

La medida más empleada en el análisis económico para cuantificar el bienestar individual es el *excedente del consumidor*. Originalmente propuesto por Dupuit (1848), se trata de una medida monetaria que permite valorar el bienestar individual alcanzado por un agente en su condición de consumidor. El excedente del consumidor ofrece, además, una cardinalización de la utilidad consistente con los axiomas de utilidad creciente y utilidad marginal decreciente aceptados con generalidad en la teoría de consumidor, siendo su simplicidad de cálculo a partir de la función de demanda y de los cambios en los precios de mercado del bien analizado su principal atractivo (Harberger, 1971). No obstante, como veremos más adelante, existen otras medidas para medir cambios en el bienestar individual más depuradas conceptualmente, que también permiten obtener estimaciones de las variaciones de bienestar individual generadas por cambios asignativos.

Para definir el concepto de excedente del consumidor partimos de la función de demanda de un determinado bien o servicio, D_X. Por ejemplo, supongamos que x es el consumo mensual de productos lácteos enriquecidos para niños, cuya demanda estimada de forma lineal para el nivel de renta medio de los hogares está representada por la siguiente función, $p_X = 20–2x$. En el momento actual, la oferta a corto plazo tiene fijado un precio de mercado de 12 euros el litro de estos productos incluidos en el bien homogéneo x. Por consiguiente, igualando la demanda al precio de oferta, el equilibrio en este mercado arroja una cantidad mensual consumida por el hogar representativo de 4 litros de x al mes.

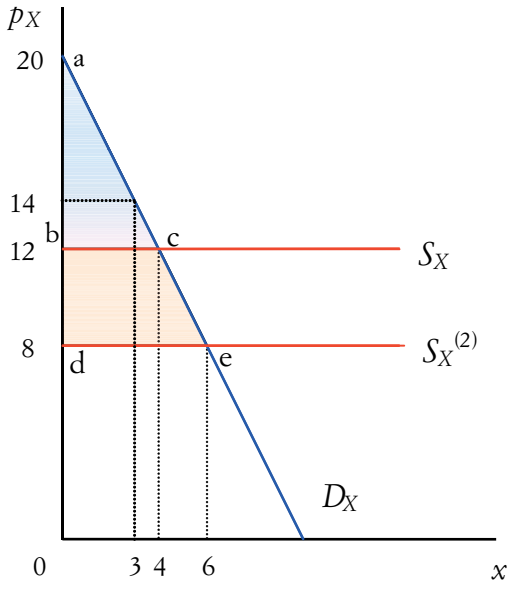


FIGURA 1

Supongamos que el gobierno aprueba una subvención dirigida a los fabricantes de este tipo de productos que reduce el precio unitario por litro de producto de 12 a 8 euros. En la Figura 1, puesto que consideramos a corto plazo el precio como dado, la oferta S_X pasa a ser $S_X^{(2)}$. En el nuevo punto de equilibrio (e) el nuevo precio de 8 euros permite un incremento del consumo mensual de x de 4 a 5 litros. Cabe ahora preguntarnos en qué medida esta política de liberalización de este mercado ha podido mejorar el bienestar de los hogares consumidores de estos productos.

Para responder a esta pregunta conviene recordar qué representa un punto de la curva de demanda de un bien. De acuerdo con el nivel de renta de un consumidor, las preferencias de éste por los bienes de consumo disponibles, y dados unos precios, la curva de demanda representa la cantidad máxima que un individuo o un hogar estaría dispuesto a pagar por cada unidad del bien x consumida. Por tanto, en el punto de equilibrio (c), el hogar representativo estaría dispuesto a pagar por el litro número cuatro de x que desean consumir el importe que determina la curva de demanda para , es decir 12 euros. Si comparamos este precio con el que estarían dispuestos a pagar por el tercer litro, 14 euros, podemos considerar, en cierto sentido, que al adquirir el litro número cuatro se está obteniendo un excedente de 2 euros por adquirir un litro más. Si retrocedemos este razonamiento desde el primer litro, el consumir 4 litros, otorga un excedente de 6 euros por el primer litro consumido (la diferencia entre los 18

euros que se estaba dispuesto a pagar por ese primer litro y los 12 que realmente se pagan al consumir 4 litros), y de 4 euros por el segundo (la diferencia entre los 16 euros que según la curva de demanda se estaba dispuesto a pagar por él y los 12 euros pagados). En total, cuando se consumen 4 litros de productos lácteos enriquecidos para niños, y el precio de mercado es 12 euros, el hogar representativo estaría obteniendo un *excedente del consumidor* de 12 euros (6 euros por el primer litro, 4 euros por el segundo litro y 2 euros por el tercero).

La reforma liberalizadora aprobada reduce el precio de mercado hasta 8 euros. Si ahora aplicamos el mismo razonamiento para construir el excedente del consumidor, el primer litro consumido aporta un excedente de 10 euros (18 que se estaba dispuesto a pagar menos los 8 del precio que se paga), el segundo 8 euros (la diferencia entre 16 y 8 euros), el tercero 6 euros (14 menos 8 euros), mientras que el cuarto litro, que antes no proporcionaba excedente, ahora aporta 4 euros (la diferencia entre la cantidad máxima que se está dispuesto a pagar por esa cuarta unidad, 12 euros, y los 8 del precio pagado). Además, como hemos visto, esta rebaja del precio de mercado hasta 8 euros supone un aumento de la cantidad demandada, que ahora llega a 6 litros. Por consiguiente, también la quinta unidad consumida también contribuye en 2 euros al excedente del consumidor, puesto que el hogar estaría dispuesto a pagar por ella 10 euros y el precio satisfecho es tan solo de 8. Por tanto, el *excedente del consumidor* tras la reducción del precio de mercado pasa a ser de 30 euros.

Ahora ya estamos en condiciones de contestar a la pregunta formulada. El bienestar individual antes de la decisión liberalizadora del gobierno se cuantificaba, a través del excedente del consumidor, en 12 euros. Tras la adopción de esta medida, la rebaja en el precio de mercado y el consiguiente aumento de la cantidad demandada ha supuesto una nueva cuantificación del bienestar del hogar de 30 euros. En consecuencia, a través de la diferencia entre el excedentes del consumidor posterior y previo al cambio asignativo, estimamos que la reforma analizada supone una mejora en el bienestar individual que se cifra en 18 euros.

Si nos fijamos en la Figura 1, vemos que el *excedente del consumidor* con el que hemos cuantificado el bienestar individual previo a la reforma liberalizadora puede ser aproximado por el área comprendida entre la curva de demanda D_X y la curva de oferta S_X que determina a corto plazo el precio de mercado de 12 euros. Es decir, puesto que la función de demanda estimada es una línea recta, el valor de este excedente del consumidor coincide con el área del triángulo  . De igual forma, el área del triángulo  aproxima el valor del excedente del consumidor que permite cuantificar el bienestar individual tras la medida del gobierno que reduce

el precio de mercado hasta 8 euros el litro. La diferencia entre ambas superficies, representada por el área del trapezoide $bcde$, constituye la medida del cambio de bienestar individual generada por la reforma aprobada.

El error de medición en el que podemos estar incurriendo al obtener la cuantificación del cambio de bienestar mediante esta diferencia de áreas depende del grado de continuidad del consumo correspondiente al bien analizado. En aquellos casos en los que el consumo de *x* sea una variable continua o cuasi-continua, el excedente del consumidor será la suma de los excedentes imputados a cada aumento infinitesimal de consumo. Por tanto, el área comprendida entre la curva de demanda y el precio de mercado será una cuantificación bastante exacta del excedente del consumidor. En cambio, si el consumo de *x* es de carácter discreto, existirá un error de cálculo al no sumarse los valores concretos de los excedentes parciales de cada unidad consumida. Con los datos recogidos en el ejemplo con el que hemos ilustrado este concepto, vemos que el valor del excedente del consumidor previo a la reforma, calculado como el área del triángulo abc es de 16 euros $(\frac{1}{2}\cdot(20-12)\times 4)$, en lugar de los 12 euros calculados de forma discreta. Igualmente, el valor del excedente posreforma calculado a través del área del triángulo adc asciende a 36 euros $(\frac{1}{2}\cdot(20-8)\times 6)$, mientras que, como vimos, cuando se computó unidad a unidad ascendía a 30 euros.

En el trabajo empírico, la obtención el excedente del consumidor se realiza a través del cálculo del área comprendida entre la curva de demanda y el precio relevante para su determinación, ya sea el precio de mercado, éste corregido por una intervención pública como el establecimiento de un impuesto o la concesión de un transferencia o subsidio, o un precio sometido a regulación.

El excedente del productor

Siguiendo el mismo razonamiento empleado para el excedente del consumidor, también podemos medir el bienestar alcanzado por los oferentes. En este caso, el concepto de *excedente del productor* (también denominado en ocasiones, *excedente del oferente*) resulta aún más intuitivo, pues como vamos a ver a continuación, esta cuantificación monetaria del bienestar individual no es otra cosa que la medición del beneficio empresarial obtenido bien por la venta de un bien o servicio, bien por la oferta un factor productivo, como sucede en la oferta de trabajo.

De igual forma que en el caso del excedente del consumidor, el *excedente del productor* se identifica con el área comprendida entre la curva de oferta del bien o factor productivo y el precio de mercado correspondiente. En la Figura 2 representamos el excedente del productor correspondiente a la oferta de trabajo. Este *excedente* es frecuentemente utilizado en el análisis de las distorsiones generadas por el impuesto sobre la renta personal en las decisiones de oferta de horas de trabajo.

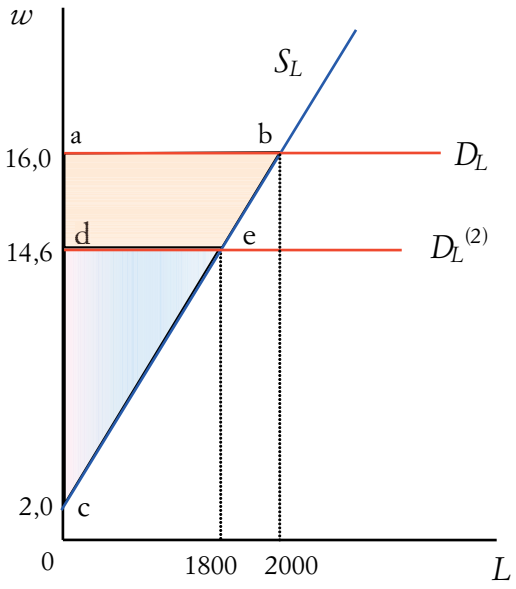


FIGURA 2

Supongamos que en la situación de partida, el trabajador representativo está dispuesto a trabajar 2000 horas anuales teniendo en cuenta que el mercado le retribuye con salario neto del impuesto sobre la renta personal de 16 euros por hora trabajada. En esta situación, el valor de su *excedente como oferente de trabajo* vendrá representado por el área comprendida entre su función de oferta de trabajo (S_L) y el precio de mercado definido por la demanda de trabajo (D_L). Por tanto, el valor monetario de este excedente será de 14.000 euros, superficie del triángulo $\overline{abc}$, $(\frac{1}{2}\cdot(16-2)\times 2000)$. Al igual que en el caso del consumidor, este excedente del productor es el resultado de agregar para todas las horas trabajadas (2000 al año) la cantidad suplementaria de renta salarial que el trabajador percibe si se tiene en cuenta cuál sería el salario unitario que se estaría dis-

puesto a aceptar por esa hora marginal [1]. La diferencia de cálculo entre esta aproximación discreta y la estimación a partir de la función de oferta dependerá, como ya indicamos, de lo pequeña que sea la unidad en la que medimos la cantidad ofertada.

En la misma figura, evaluamos el impacto que tiene sobre el bienestar del trabajador el aumento del impuesto sobre la renta personal. Supongamos que tras esta reforma, el nuevo impuesto reduce el salario por hora en el mercado hasta 14,6 euros (desde los 16 anteriores). En respuesta a este cambio, el comportamiento del trabajador, representado por su curva de oferta de trabajo, le llevará a reducir la cantidad de horas anuales ofertadas, de 2000 a 1800. Por tanto, el nuevo excedente del oferente pasará a ser el área del triángulo $\overline{adc}$, definido ahora por el nuevo salario unitario fijado por el mercado a través de $D_L^{(2)}$, es decir, 11340 euros $\{\frac{1}{2}\cdot[14{,}6-2]\times 1800\}$. De nuevo, la comparación entre ambos excedentes nos ofrece una medida monetaria del cambio de bienestar al que se enfrenta el trabajador tras esta reforma impositiva considerada: su bienestar se reduce en 2660 euros, al pasar de 14000 a 11340 euros.

Debemos prestar atención al potencial explicativo de esta medida de cambio de bienestar. En principio, podríamos pensar que la pérdida de bienestar de una reforma impositiva como la sugerida sería mayor, pues el trabajador al ver reducido su salario unitario neto de 16 a 14,6 euros, estaría viendo reducidos sus ingresos salariales netos en 2800 euros (1,4 euros × 2000 horas). Sin embargo, en la Figura 2 observamos que existe una respuesta por parte del trabajador, que ante este cambio en el precio del factor reduce en 200 la cantidad de horas ofertadas. En la medida que aceptamos que la oferta de trabajo es consecuencia de la optimización de su utilidad ante la decisión ocio–consumo para todo el rango de salarios unitarios factibles, su comportamiento muestra la existencia de un *efecto sustitución* que lleva al trabajador a sustituir consumo (renta) por ocio, ante el abaratamiento del coste de oportunidad de este último [2]. Tras la reforma, renunciar a una hora adicional de ocio aporta 1,4 euros menos de capacidad de consumo, lo que en términos relativos hace más atractiva esa hora de ocio. De este modo, compensa una parte de la pérdida potencial de bienestar que le supone la reducción de consumo (renta) (2800 euros) con el aumento de bienestar generado por el incremento de 200 horas de ocio (140 euros). Por tanto, su reducción de bienestar queda fijada en los

[1] La intersección de la curva de oferta de trabajo en el eje del salario para un valor de 2 euros representa el salario de reserva por debajo del cual el trabajador no estaría dispuesto a incorporarse al mercado de trabajo.

[2] Alternativamente, un trabajador podría mostrar un efecto renta predominante si sus preferencias le llevaran a aumentar la cantidad de horas ofertadas para recuperar el nivel salarial previo al aumento del impuesto.

2660 euros estimados por la diferencia entre los excedentes del oferente previo y posterior a la subida del impuesto.

Distorsiones impositivas y coste de bienestar

Una aplicación frecuente de estas medidas de cambio de bienestar basadas en los excedentes del consumidor o del productor es su utilización para cuantificar la distorsión generada por la introducción de un impuesto o por la modificación de uno ya existente. Volvamos a la reforma impositiva planteada en la Figura 2. Tras el aumento del impuesto que reduce el salario neto percibido de 16 a 14,6 euros, vemos que la recaudación anual aumentará por trabajador en 2520 euros (1,4 euros por 1800 horas trabajadas tras este aumento del impuesto sobre la renta personal). Como vemos en la Figura 3, este importe es parte del cambio de bienestar que hemos medido: el área correspondiente al rectángulo [redacted]. En sentido estricto, no puede decirse que este aumento de la recaudación suponga desde un punto de vista asignativo una pérdida de bienestar, sino un cambio de propiedad para estos recursos que ahora han pasado a manos del gobierno (transitoriamente, hasta que sean utilizados en programas de gasto que influirán a su vez en el bienestar individual).

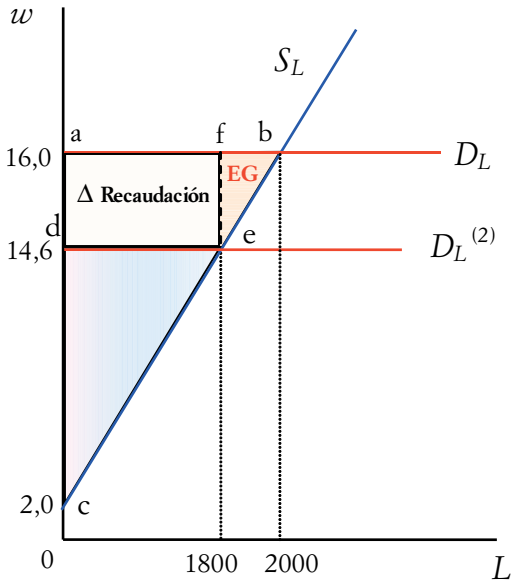


FIGURA 3

Sin embargo, vemos que una vez descontado el importe de la recaudación (los 2550 euros), aún resta una parte del cambio de bienestar provocado por el aumento impositivo, el área del triángulo EG. Esta parte de la pérdida de excedente es el *coste de eficiencia* generado por esta reforma impositiva, denominado *exceso de gravamen* (EG). Se trata de la valoración monetaria, 140 euros $\frac{1}{2}\cdot(16-14{,}6)\times(2000-1800)$, del bienestar que pierde el trabajador por su respuesta de reducción de las horas ofertadas, teniendo en cuenta que su elección óptima entre consumo y ocio antes del aumento del impuesto le llevaba a preferir una combinación con 2000 horas anuales de trabajo (32000 euros de renta disponible), frente a la resultante de 1800 horas anuales de trabajo (26280 euros). Podemos comprobar que esta coste de bienestar coincide, precisamente, con la diferencia entre los 2800 euros potenciales de recaudación que se pierden por al no trabajarse 200 horas (1,4 euros × 200 horas) y los 2660 euros en los que se ha cuantificado el cambio en el bienestar tras la reforma.

Este exceso de gravamen se puede calcular de forma directa a través de la siguiente expresión, con el requisito de que la forma funcional estimada para la oferta de trabajo sea una recta:

$$EG=\frac{1}{2}\left|\varepsilon_L^s\right|\cdot L_0\cdot w_0\cdot t_L^2 \qquad [1]$$

donde ε_L^s es la elasticidad de la oferta de trabajo, w_0 y L_0 son, respectivamente, el salario neto y la cantidad de horas ofertadas antes del cambio evaluado, y t_L el tipo de gravamen del impuesto. Esta expresión del exceso de gravamen resulta muy interesante, pues nos permite analizar la sensibilidad del coste de eficiencia introducido por el impuesto a la elasticidad de la oferta de trabajo o a la intensidad del cambio en el impuesto. En nuestro ejemplo, el tipo de gravamen introducido por la reforma considerada es del 8,75% $(16\times(1-0{,}0875)=14{,}6)$, y la elasticidad de la oferta de trabajo, 1,142857 $(2000-1800)/2000/(16-14{,}6)/16)$. Por consiguiente, obtenemos el mismo valor del exceso de gravamen que mediante la descomposición del cambio en bienestar mostrado en la Figura 3:

$$EG=\frac{1}{2}\left|1{,}142857\right|\cdot 2000\cdot 16\cdot 0{,}0875^2=140 \text{ euros}$$

Para el trabajo aplicado hay que tener en cuenta la información disponible. Normalmente, el análisis empírico de las distorsiones impositivas (también las generadas por subvenciones u otras medidas de intervención que afecten a los precios de mercado de bienes o factores)

suele realizarse sobre los datos observados tras la adopción de una reforma. En este caso, no suele ser factible contar con información sobre las horas inicialmente trabajadas (o las cantidades inicialmente consumidas de un bien, para un impuesto sobre el consumo) ni con información sobre el salario unitario previo a la reforma (o el precio de mercado anterior a la modificación impositiva sobre el consumo). En estos casos, se suele utilizar la aproximación al cálculo del exceso de gravamen basada en el equilibrio observado en el punto "e" en lugar del punto "b" (de la Figura 3), donde hay que tener en cuenta que debe ser evaluada en "e":

$$EG = \frac{1}{2}\left|\varepsilon_L^d\right| \cdot L_1 \cdot w_1 \cdot \left(\frac{t_z}{1-t_z}\right)^2 \qquad [2]$$

Como podemos ver, considerando que la elasticidad en ambos puntos es idéntica, se trata de una buena aproximación, aunque como tal, ofrece una cuantificación distinta a la obtenida a partir del equilibrio previo a la reforma:

$$EG = \frac{1}{2}\left|1{,}142857\right| \cdot 1800 \cdot 14{,}6 \cdot \left(\frac{0{,}0875}{1-0{,}0875}\right)^2 = 138{,}08 \text{ euros}$$

EJERCICIO PROPUESTO N.º 1

Evaluar el cambio en el bienestar individual experimentado por los consumidores que se enfrentan a la introducción de un impuesto sobre el consumo de servicios de alojamiento hotelero del 5% del precio de venta (antes de impuestos). La función de demanda estimada de servicios hoteleros (para un estándar medio de habitación en el sector hotelero español) es $p_H = 400 - 20H$. A corto plazo, el precio de mercado por habitación/día, antes de la introducción de este impuesto, es de 100 euros. También se desea conocer cual será el exceso de gravamen introducido por el nuevo impuesto, así como la cantidad que se estima recaudar.

Otras medidas para el análisis de los cambios de bienestar individual

Además del concepto de excedente del consumidor (o, en su caso, del productor) existen otras medidas monetarias para evaluar los cambios de bienestar producidos por alteraciones en los precios, bien por el propio comportamiento del mercado, bien por las diversas intervenciones del sector público.

Posiblemente las dos más utilizadas sean la *variación equivalente* y la *variación compensatoria*, ambas propuestas por Hicks (1939), a partir de la noción *marshalliana* del excedente del consumidor. Como señala Hausman (1981), estas dos medidas, compatibles con las nociones de compensación empleadas en la Economía del bienestar, suponen un refinamiento teórico notable respecto de la utilización pionera de los excedentes del consumidor y el productor. En relación con su utilización empírica, Willig (1976) ha evaluado los errores de medición que pueden cometerse al emplear estos excedentes basados en las curvas ordinarias (*marshallianas*) de demanda u oferta, en lugar de las correspondientes medidas *hicksianas* construidas a partir de las respectivas curvas compensadas, encontrando que en un buen número de casos la utilización de los primeros ofrece una buena aproximación, lo que ayudaría a comprender la generalidad de su uso. La variación compensatoria y la variación equivalente difieren en el criterio de compensación subyacente para su cómputo.

La variación compensatoria

En el caso de la *variación compensatoria*, la cuantificación monetaria del cambio en bienestar se obtiene bajo el supuesto de que tras la modificación del precio que afecta al consumidor (o en su caso al oferente), éste debe ver ajustado su nivel de renta para poder mantener el nivel de utilidad que alcanzaba antes de esa alteración del precio. Por consiguiente, esta medida se identifica con la cantidad de euros que debe recibir un agente afectado para que con los nuevos precios pueda disfrutar del mismo nivel de bienestar que tenía antes del cambio. Desde un punto de vista ético-normativo, este criterio supone que el consumidor ha de recibir una cantidad de dinero adicional para compensarle de la pérdida de utilidad que le general la variación del precio.

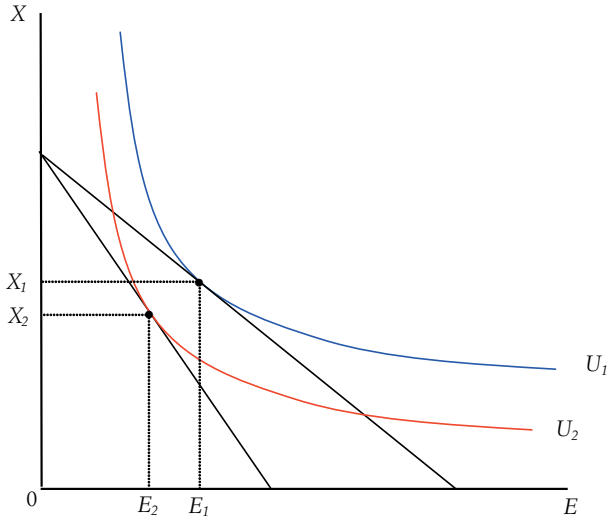


FIGURA 4

Para ilustrar esta medida planteamos el siguiente ejemplo. En la figura 4 recogemos los niveles de utilidad alcanzados por un hogar representativo en relación con su consumo anual de electricidad (E) antes (1) y después (2) de la introducción de un impuesto especial sobre el consumo de energía, t_E del 20% del precio antes de impuestos. Tras la introducción del impuesto, la restricción presupuestaria del hogar, $R = E_1 \cdot p_1^E + X_1 \cdot p_1^X$ (donde X representa el resto de bienes de la cesta de consumo, a un precio homogéneo p_X), pasa a ser $R = E_2 \cdot p_1^E + X_2 \cdot p_1^X$. Como vemos, tanto el nivel de renta monetaria R como el precio del resto de bienes p_1^X no se ven afectados, aunque sí, dadas las preferencias del hogar, la cantidad consumida de ambos bienes, E_2 y X_2.

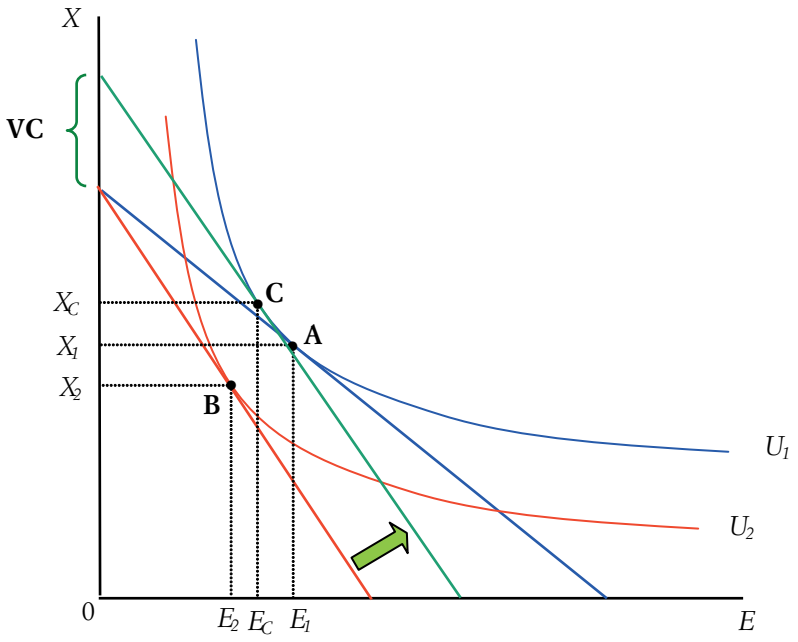


FIGURA 5

De acuerdo con la definición que hemos establecido para la *variación compensatoria*, en la Figura 5 buscamos el incremento de renta que habría que proporcionar a este hogar para que con el nuevo precio de la electricidad (incluido el impuesto especial introducido) no se viese mermado su nivel de utilidad. En la medida que el cambio en el precio de la electricidad, $p_1^E \mapsto p_1^E \cdot (1 + t_E) = p_2^E$, supone un cambio en la pendiente de la res-

tricción presupuestaria, que pasa de p_1^E/p_1^X a p_2^E/p_1^X, para encontrar este nivel de renta que permitiría al hogar mantener el nivel de utilidad previo debemos desplazar la nueva recta presupuestaria $R=E_2\cdot p_2^E+X_2\cdot p_1^X$ en paralelo (manteniendo su pendiente) hasta encontrar un nuevo punto de tangencia con la curva de indiferencia que proporciona el nivel de utilidad U_1. Este nuevo equilibrio determinaría las cantidades tanto de energía eléctrica como del resto de bienes que podrían consumirse teniendo en cuenta el nuevo precio de E, p_2^E, si se dispusiese de un nivel de renta R^C que lo permitiese. Si llamamos R_C a este nivel de renta incrementado, $R_C=E_C\cdot p_2^E+X_C\cdot p_1^X$ será la ecuación de la correspondiente recta presupuestaria, paralela como vemos a la recta presupuestaria $R=E_2\cdot p_2^E+X_2\cdot p_1^X$. En consecuencia, el cambio de bienestar medido a través de la *variación compensatoria* (*VC*) se puede calcular como:

$$VC=R_C-R=p_2^E\cdot(E_C-E_2)+p_1^X\cdot(X_C-X_2) \qquad [3]$$

La variación equivalente

La *variación equivalente*, por su parte, fija como referencia la utilidad alcanzada tras la modificación del precio. Por tanto, identificamos esta medida de cambio de bienestar con la cantidad de renta a la que debería renunciar el hogar afectado para que con los precios originales se sitúe en el mismo nivel de utilidad alcanzado realmente tras el cambio. Desde una perspectiva ético-normativa, la *variación equivalente* puede explicarse como la cantidad de renta que un agente estaría dispuesto a pagar por evitar la pérdida de utilidad (de bienestar) que le genera un aumento del precio [3].

En la figura 6 ilustramos el cómputo de la variación equivalente para el ejemplo planteado. Puesto que ahora las referencias son el nivel de utilidad que se obtenido por el hogar tras la modificación del precio de E y los precios originales y p_1^X y p_1^E, esta medida de cambio de bienestar puede calcularse desplazando en paralelo la restricción presupuestaria inicial $R=E_1\cdot p_1^E+X_1\cdot p_1^X$ hasta encontrar su punto de tangencia con la curva de

[3] Si la modificación del precio supusiese su reducción (p.e. por la introducción de una subvención o por la apertura a la competencia de ese mercado), podríamos explicar la variación equivalente como la compensación que exigiría ese hogar por no disfrutar del aumento de utilidad (de bienestar) que le reportaría la medida adoptada.

indiferencia que proporciona ese nivel de utilidad U_2. Si tenemos en cuenta que la ecuación de esta nueva recta presupuestaria será $R_{\blacksquare} = E_{\blacksquare} \cdot p_1^E + X_{\blacksquare} \cdot p_1^X$, el cambio de bienestar medido a través de la *variación equivalente* (*VE*) puede calcularse como:

$$VE = R - R_{\blacksquare} = p_1^E \cdot (E_1 - E_{\blacksquare}) + p_1^X \cdot (X_1 - X_{\blacksquare}) \qquad [4]$$

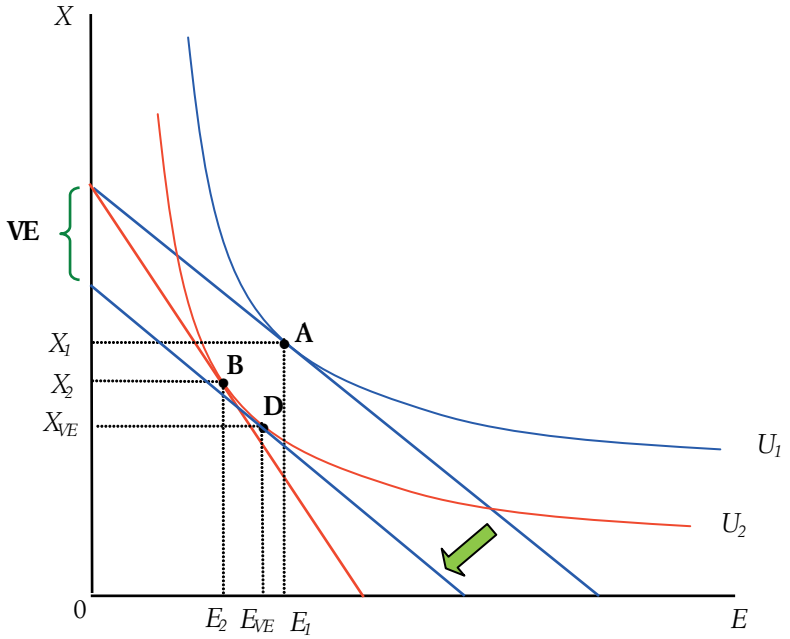


FIGURA 6

El cálculo de las variaciones compensatoria y equivalente a partir de las curvas de demanda/oferta compensadas

Las nociones que están detrás de las dos medidas de cambio de bienestar individual que acabamos de presentar están relacionadas, respectivamente, con las dos definiciones tradicionales del efecto sustitución empleadas en Microeconomía, el *efecto sustitución* de Hicks y el *efecto sustitución* de Slutsky.

En el primer caso, el *efecto sustitución* de Hicks recoge la variación en la cantidad demandada de un bien (o la cantidad ofertada de un factor,

en su caso) al modificarse su precio, bajo la condición de que el nivel de utilidad del consumidor permanece constante. Si nos fijamos en la Figura 5, esta respuesta es la que se produce al comparar los equilibrios A y C: al elevarse el precio de la electricidad a p_1^E a p_2^E, la cantidad consumida se reduce de E_1 a E_C.. De acuerdo con la forma en la que se ha obtenido el equilibrio C en la Figura 5, con esta modificación de la cantidad demandada (y para un nivel de consumo X_C del resto de bienes) se consigue mantener al consumidor en su nivel de utilidad inicial (U_1), aunque obviamente con una alteración del cociente de precios relativos (reflejada en el cambio de pendiente de la restricción presupuestaria).

Si relacionamos los posibles cambios en los precios con las respuestas en la cantidad demandada (efecto sustitución de Hicks) estamos en condiciones de obtener una curva de demanda en la que el nivel de utilidad del consumidor se mantiene constante. Puesto que para mantener la utilidad ante los cambios en el precio es necesario compensar al consumidor, esta curva de demanda *hicksiana* (o, en su caso, de oferta) es denominada curva de demanda compensada.

Alternativamente, el efecto sustitución generado por la alteración de los precios puede medirse bajo la condición de que el poder adquisitivo del consumidor se mantiene constante, en lugar de suponer que un nivel de utilidad fijo. Si observamos en la Figura 6, si el cociente de precios relativos permaneciese constante, el desplazamiento paralelo de la restricción presupuestaria inicial hasta alcanzar el equilibrio D (tangencia con la curva de indiferencia U_2) representa la pérdida de poder adquisitivo generada por el aumento del precio de E de p_1^E a p_2^E. Por tanto, la reducción en la cantidad demandada desde E_1 a E_{VE}, consecuencia del movimiento del equilibrio inicial A al equilibrio D es la respuesta correspondiente a este *efecto sustitución de Slutsky*. De igual manera que en el otro efecto sustitución, si relacionamos estos cambios en la cantidad demandada con los cambios en el precio que la generan obtenemos otra variante de la curva de demanda compensada, la denominada *curva de demanda de Slutsky*.

En la Figura 7 presentamos la relación existente entre la curva de demanda ordinaria o *marshalliana* (CDO) y estas dos curvas de demanda compensadas construidas a partir de las dos definiciones alternativas del efecto sustitución. Tanto la curva de demanda compensada de Hicks (CDH) como la de Slutsky (CDS) siempre tienen pendiente negativa (no resulta difícil comprobar que esto es debido, precisamente, al requisito de compensación establecido), mientras que en el caso de la demanda ordinaria teóricamente es posible encontrar una pendiente positiva si estamos ante un bien inferior. Debemos destacar que cuando las variaciones en los precios consideradas son pequeñas, ambos efectos sustitución son prácticamente idénticos.

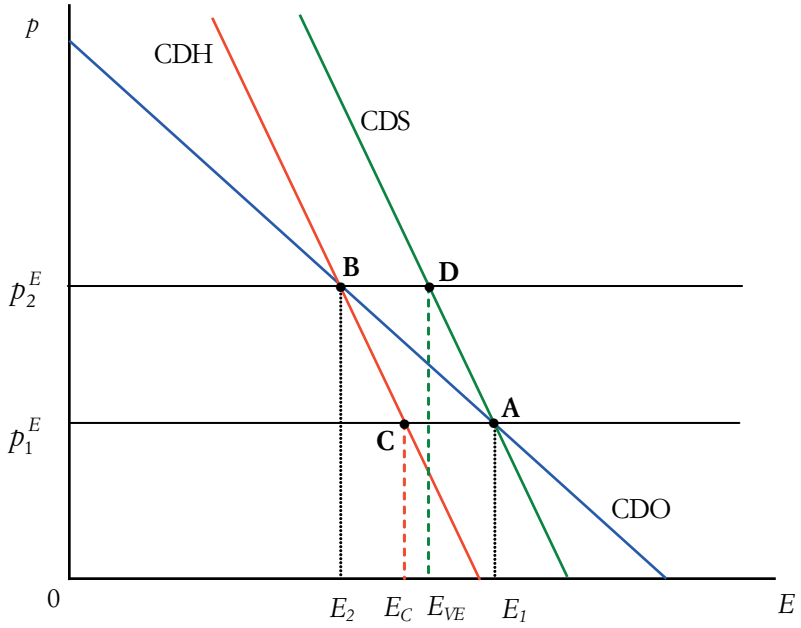


FIGURA 7

A partir de estas dos curvas de demanda compensada podemos computar como medidas de área tanto el valor de la variación compensatoria como el de la variación equivalente. El procedimiento es análogo al seguido en el caso del excedente del consumidor, con la diferencia que ahora las áreas son delimitadas por las correspondientes curvas compensadas en lugar de por la demanda ordinaria. Puesto que la curva CDH ha sido obtenida bajo el criterio de compensación que mantiene al consumidor en el nivel de utilidad inicial, el cambio de bienestar medido a través de la variación compensatoria se corresponde con el área $\overline{p_1^E BCp_2^E}$, que sombreamos en rojo en la Figura 8a. De otra forma, el cambio de bienestar medido por la variación compensatoria puede computarse a través de la siguiente expresión que incorpora tanto el cambio en el precio como la respuesta del efecto sustitución de Hicks:

$$VC=\left(p_1^E-p_2^E\right)\cdot\left[E_2+\tfrac{1}{2}\cdot\left(E_C-E_2\right)\right] \qquad [5]$$

Por su parte, el cambio de bienestar medido a través de la variación equivalente viene determinado por la curva CDS, correspondiendo su valor monetario al área $\overline{p_1^E DAp_2^E}$, sombreada en verde en la Figura 8b. Al igual que para la variación compensatoria, el cómputo de la variación equi-

valente también puede obtenerse a partir del cambio en los precios y la respuesta en la cantidad demandada correspondiente al efecto sustitución de Slutsky:

$$VE=(p_2^E-p_1^E)\cdot[E_{m'}+k_1\cdot(E_1-E_{m'})] \qquad [6]$$

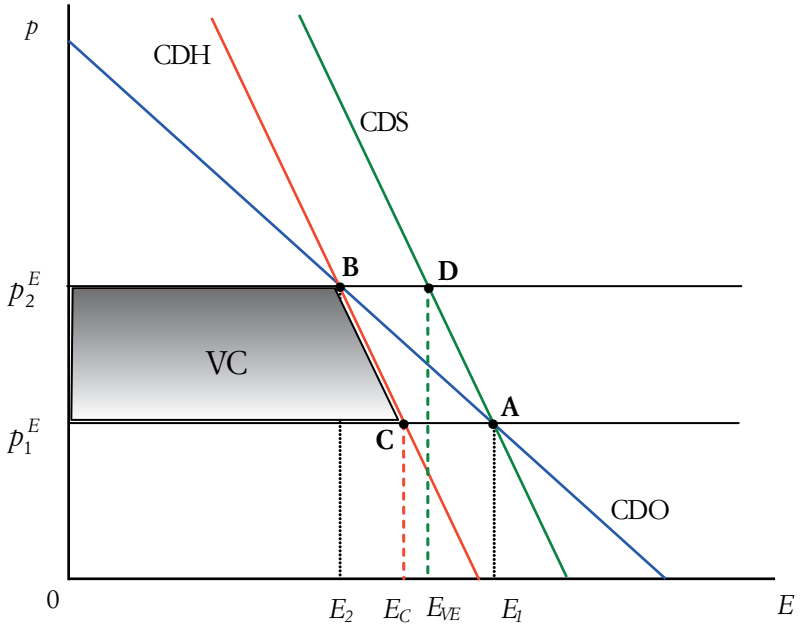


FIGURA 8

En ambas figuras también podemos identificar el cambio de bienestar en términos de excedente del consumidor, a través del área $\overline{p_1^E BAp_1^E}$ delimitado por la curva de demanda ordinaria (CDO). De igual modo que para las variaciones de bienestar *hicksianas*, el valor monetario de la variación producida en el excedente del consumidor (VEC) se puede obtener a partir del cambio en los precios y, en este caso, de la respuesta de la cantidad demandada determinada conjuntamente por el efecto renta y el efecto sustitución:

$$VEC=(p_2^E-p_1^E)\cdot[E_1+k_1\cdot(E_1-E_2)] \qquad [7]$$

Comparando estas tres áreas observamos cómo a pesar de que estamos evaluando un único cambio en precios —el generado por la introduc-

ción del impuesto sobre el consumo de electricidad— disponemos de tres mediciones distintas del cambio de bienestar individual. Esto es debido a que las curvas de demanda utilizadas para delimitar estas áreas son distintas en los tres casos. En el caso del excedente del consumidor, no existe criterio de compensación, por lo que se utiliza la curva de demanda ordinaria. En las otras dos medidas, la curva de demanda compensada utilizada en cada caso es distinta, al diferir como vimos el criterio de compensación empleado para su obtención.

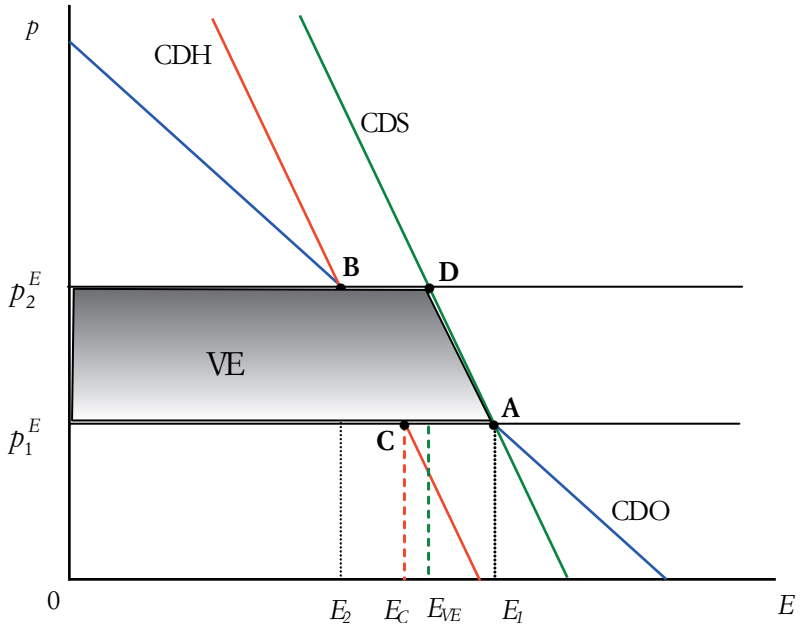


FIGURA 8b

Si comparamos ambas áreas con la proporcionada por el excedente del consumidor (EC), se verifica la siguiente ordenación entre estas tres medidas de cambio de bienestar individual [4]:

$$VC < EC < VE \qquad [8]$$

No obstante, hay que reseñar un interesante resultado desde el punto de vista del análisis microeconómico. La variación compensatoria y la variación equivalente ofrecen la misma cuantificación del cambio en bien-

[4] Willig (1976) demuestra esta relación, así como que la diferencia entre la variación compensatoria y la variación equivalente será mayor cuanto mayor sea la elasticidad renta de la demanda, y menor cuanto más se acerque la elasticidad precio de la demanda a la unidad.

estar cuando la función de utilidad es cuasilineal. Bajo esta condición, las curvas de indiferencia son paralelas, por lo que su distancia es independiente del punto de tangencia con la recta presupuestaria. En este caso, además, puede comprobar el lector que estas medidas también coinciden con la proporcionada al utilizar el excedente del consumidor.

A continuación proponemos el siguiente ejercicio de evaluación en el que podemos realizar una interesante comparación entre las tres medidas de cambio de bienestar propuestas, destacando en cada caso sus ventajas y limitaciones.

EJERCICIO PROPUESTO N.º 2

Evaluar el cambio en el bienestar individual experimentado por los consumidores que se enfrentan a la introducción de un impuesto sobre el consumo de servicios de alojamiento hotelero del 5% del precio de venta (antes de impuestos). La función de demanda estimada de servicios hoteleros (para un estándar medio de habitación en el sector hotelero español) es $p_H = 400 - 20H$. A corto plazo, el precio de mercado por habitación/día, antes de la introducción de este impuesto, es de 100 euros. También se desea conocer cual será el exceso de gravamen introducido por el nuevo impuesto, así como la cantidad que se estima recaudar.

8.3. EVALUANDO EL BIENESTAR SOCIAL

¿Debe el análisis económico incorporar en la evaluación de las políticas públicas sus consecuencias distributivas? Si nos atenemos a la atención prestada por la literatura especializada en las últimas dos décadas, lo mismo que si atendemos a las preocupaciones mostradas por la sociedad, la respuesta debe ser afirmativa. Como señala Rosen (2005), la Economía del Bienestar pone de manifiesto en sus postulados que la eficiencia, por sí misma, no es suficiente para evaluar una situación económica, siendo necesario acudir a otros criterios distintos. Desde una perspectiva colectiva, resulta complicado sostener que la eficiencia sea por si sola un criterio suficiente para establecer un orden de preferencia entre asignaciones alternativas. Por supuesto que podría defenderse que la eficiencia es el único criterio de evaluación relevante, pero también debería admitirse que esta posición, en sí misma, no es más que un juicio de valor.

No obstante, si aceptamos la importancia de los aspectos distributivos a la hora de valorar un resultado asignativo, tampoco podemos pasar

por alto la necesidad de juzgar entre alternativas que al menos proporcionen un valor económico aceptable, o lo que es lo mismo, que no supongan una infrautilización o un despilfarro de recursos. En este sentido, parece lógico exigir que una ordenación valorativa de asignaciones en relación con el nivel de bienestar que aportan a la sociedad —bienestar social— valore conjuntamente tanto el tamaño del pastel disponible para los ciudadanos como el grado de igualdad con el que éste se distribuye. De hecho, si no se tuviese en cuenta el principio de eficiencia (del que depende el tamaño del pastel), podríamos llegar a considerar que el bienestar de los habitantes de un país es superior simplemente si la igualdad con la que se distribuye su renta es menor, sin tener en cuenta su renta media. De hecho, en algunos casos, países con una renta *per cápita* muy reducida muestran niveles de desigualdad inferiores a los existentes en países desarrollados, sin que parezca razonable argumentar que los habitantes de aquellos gozan de un nivel de bienestar social superior al alcanzado en estos otros. Esta misma situación puede plantearse desde un punto de vista regional cuando encontramos dentro de un mismo país regiones ricas con una renta *per cápita* bastante notablemente superior a la renta media nacional pero distribuida con una desigualdad elevada, junto a regiones pobres en las que una reducida renta *per cápita* se distribuye más igualitariamente.

De acuerdo con estas consideraciones, la evaluación del bienestar social consiste, en buena medida, en establecer un orden de preferencias respecto de la distribución del bienestar individual. Como hemos visto, desde la perspectiva individual, el bienestar de un ciudadano o de un hogar depende exclusivamente de sus niveles de renta o consumo y de sus preferencias, sin tener en cuenta las valoraciones del resto de individuos u hogares. En cambio, la fijación de este orden de preferencias social exige incorporar a la valoración el punto de vista de la sociedad acerca del nivel de renta o consumo de todos los ciudadanos y de su situación relativa.

Las funciones de bienestar social

La especificación de funciones de bienestar social (FBS) es la forma convencional de establecer la posición de la sociedad a la hora de valorar el bienestar social del que disfrutan sus miembros en relación con su nivel de renta o acceso al consumo. Una función de bienestar social es una expresión matemática encargada de recoger la forma concreta en la que el bienestar de la sociedad se ve influido por el bienestar individual de cada uno de sus miembros. La manera más habitual de establecer esta relación funcional es definir el bienestar social a partir de las utilidades de cada individuo i (o, en su caso, hogar) en relación con su renta o su nivel de consumo ($U(x_i)$):

$$W = F(U(x_1), U(x_2), \ldots, U(x_N)) \qquad [9]$$

Estas funciones de bienestar social, denominadas *individualistas*, poseen un atractivo criterio normativo en relación con la valoración colectiva del bienestar, como es que los individuos son considerados en las mismas como los mejores jueces de su propio bienestar.

Una condición habitualmente exigida para cualquier función de bienestar social es que su valor (W) aumente ante un incremento en la utilidad de cualquier individuo u hogar $(\partial W/\partial U(x_i) > 0; \forall i)$. Este requisito equivale a aceptar que cuando uno cualquiera de los miembros de la sociedad mejora su bienestar, la sociedad en su conjunto mejora.

Supongamos ahora la siguiente especificación concreta de la FBS genérica recogida en [9]:

$$W = \frac{1}{N} \cdot \left[U(x_1) + U(x_2) + \ldots + U(x_N) \right] \qquad [10]$$

En [10], el valor del bienestar social se identifica con el promedio de las utilidades alcanzadas por los N miembros que integran la sociedad. Por consiguiente, ante un incremento de la renta de cualquier individuo (Δx_i), el valor de la *utilidad media* (ΔW) aumentará. La única condición exigida es que los individuos muestren una preferencia positiva hacia la renta (o, en su caso, su aplicación al consumo), es decir, que la primera derivada de la función de utilidad sea positiva $\partial U(x)/\partial x > 0)$.

De acuerdo con lo expuesto, a la hora de llevar a cabo cualquier evaluación normativa del bienestar social resulta fundamental explicitar la postura ética asumida por la sociedad en relación con la desigualdad. En el ámbito de la Economía del bienestar, este criterio ético-normativo se recoge a través de la noción de *aversión a la desigualdad*. Intuitivamente, la aversión a la desigualdad representa el coste en el que la sociedad valora la «molestia» que le supone la existencia de una distribución de la renta desigual. Como señalan Kay y King (1984), este *coste de desigualdad* puede medirse a través de la reducción de la renta total que la sociedad estaría dispuesta a asumir a cambio de conseguir una distribución totalmente igualitaria.

Centrándonos en la especificación de la función de bienestar social, la aversión a la desigualdad es un concepto identificable con su concavidad, puesto que para un mismo nivel de renta total esta propiedad asegura mayores valores de W a medida que la distribución se acerca a la completa igualdad.

¿Cómo introducir este principio de *aversión a la desigualdad* en las funciones de bienestar social *individualistas*? La manera habitualmente empleada en la literatura es la propuesta por Atkinson (1970), basada en el reconocimiento del supuesto de decrecimiento de la utilidad marginal de la renta. Como es bien sabido, se trata de un supuesto tradicional en el análisis microeconómico, de aceptación prácticamente general. Parece poco discutible que el aumento en la utilidad generado por un incremento en la renta de 100 euros será mayor para un individuo que gana 1000 euros mensuales que para uno que gana 3000 euros. Igualmente, la aceptación de este supuesto es plenamente consistente con el principio de utilidad marginal decreciente en el consumo. Si admitimos que la utilidad marginal que proporciona a cualquier individuo su renta (o alternativamente su aplicación al consumo) es decreciente, $(\partial U'(x)/\partial x^2 < 0)$, aplicando la regla de derivación en cadena, *W* será también una función *cóncava*.

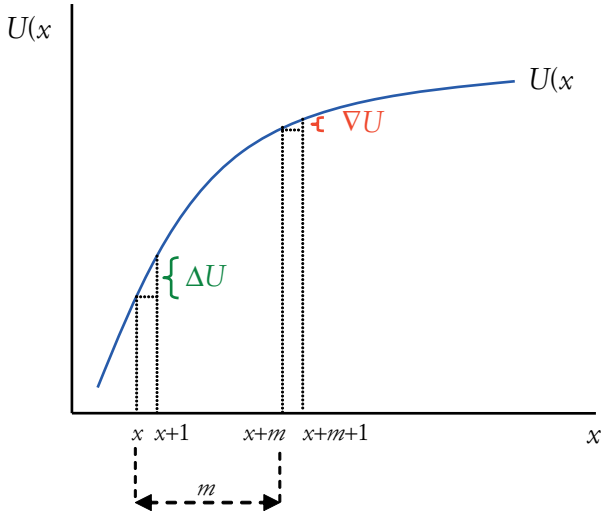


FIGURA 9

La *concavidad* de la función de bienestar social garantiza una valoración favorable al reparto más igualitario de un determinado nivel de renta total. Detrás de este resultado matemático, bastante intuitivo, encontramos el sustento ético que acabamos de exigir para este tipo de valoraciones normativas: el principio de *aversión a la desigualdad*. Además, la concavidad de la FBS supone implícitamente la aceptación del *Principio de las Transferencias* de Pigou y Dalton. Como podemos observar en la Figura 9, si y sólo si la función de utilidad $U(x)$ es estrictamente cóncava en todo el rango de rentas (o consumos), la reducción de utilidad soportada por un in-

dividuo al transferir un euro será más que compensada por el aumento de utilidad que supondrá la percepción de ese euro por cualquier individuo que posea un nivel inferior de renta (en la figura, existe una diferencia de *m* euros de renta). En consecuencia, la sucesión de transferencias «de ricos a pobres» necesarias para reducir la desigualdad hasta un determinado grado, incluso hasta alcanzar la completa igualdad, proporcionará incrementos en el bienestar social, en la medida que el valor agregado de las pérdidas de utilidad soportadas por aquellos individuos que ven reducida su renta será inferior al valor agregado de las ganancias de utilidad obtenidas por los individuos que reciben esas transferencias $(\Sigma \Delta U > \Sigma \nabla U)$.

El *Principio de las Transferencias Decrecientes* supone un criterio ético más exigente (más restrictivo) a la hora de establecer la posición del evaluador social. La FBS estará a favor de este principio si además de aceptar la estricta concavidad de la función de utilidad de los individuos (aceptación del *Principio de Transferencias*), las funciones de utilidad presentan una tercera derivada respecto de la renta (o, en su caso, del consumo) positiva $(\partial^3 U/\partial x^3 > 0)$. En este caso, la evaluación de bienestar social concederá mayor valor a una transferencia entre individuos situados a una determinada distancia de renta si ésta tiene lugar en niveles más bajos de renta, que si la misma se produce en los niveles más elevados.

De acuerdo con lo expuesto, el *Principio de Transferencias* permitiría justificar la actuación redistributiva de los gobiernos, siempre que los resultados de sus políticas reduzcan la desigualdad de la renta. Sin embargo, aunque por regla general en la evaluación de programas públicos el análisis de los efectos redistributivos se realiza de forma independiente respecto del análisis de eficiencia, no podemos obviar que las intervenciones redistributivas inciden sobre los resultados asignativos que determinan el tamaño del pastel a distribuir. Como señala Rosen (2002), una sociedad que aspire a hacer máxima de la suma de las utilidades de sus ciudadanos debe enfrentarse a un dilema inevitable: al avanzar en la igualación de la distribución de la renta, los incentivos (desincentivos) que guían las elecciones de los individuos entre ocio y renta provocarán una reducción de la cantidad total de renta disponible. Por consiguiente, el problema de la redistribución aparece replanteado en términos de cuál debe ser la distribución óptima de la renta, De otro modo, esta búsqueda obliga a considerar los costes (en términos de pérdida de renta real) que la sociedad está dispuesta a asumir para conseguir los beneficios de una mayor igualdad, lo que supone aceptar la «deseabilidad» de un cierto grado de desigualdad. Aunque la investigación en este campo no se encuentra aún demasiado desarrollada, en la actualidad disponemos de algunos interesantes resultados en el ámbito del diseño de los impuestos progresivos sobre la renta personal y su relación con la oferta de trabajo que genera esa renta (ver Onrubia, Salas y Sanz, 2005).

La disponibilidad de microdatos con información sobre la renta de los individuos o de los hogares en los que estos se integran, o sobre sus consumos, permite realizar evaluaciones de bienestar social a través de la especificación de FBS individualistas como las presentadas. Para ilustrar este tipo de aplicaciones, proponemos el siguiente ejercicio que emplea la distribución de renta incluida en el CD proporcionado.

EJERCICIO PROPUESTO N.º 3 (SOLUCIONADO)

En la sociedad cuya distribución de la renta se ofrece (Fichero XXX), el gobierno está planteando introducir una ayuda familiar por hijo en edad escolar de 500 euros anuales con el objetivo de reducir el coste educativo al que se enfrentan las familias. Alternativamente se está considerando la aplicación del programa únicamente para aquellas familias con ingresos iguales o inferiores a 20.000 euros, frente a su extensión a la totalidad de los hogares. Si suponemos que el gobierno representa a una sociedad que muestra aversión hacia la desigualdad, se desea conocer cuál será el impacto sobre el bienestar social del programa de gasto, en sus dos alternativas barajadas. Para obtener los resultados, considere una función de utilidad de la renta familiar disponible, tal que $U = x^{\frac{1}{4}}$.

En primer lugar, puesto que partimos del fichero de microdatos de los individuos procedemos a agregar la información de cada hogar familiar. Para ello, empleamos la herramienta «Informe de tablas y gráficos dinámicos» de Excel. El primer paso es marcar el bloque que contiene todos los datos en la hoja de cálculo (A1:E10173, incluyendo por tanto los encabezamientos de las variables). El segundo paso, como se muestra en la ventana, es obtener del menú de Datos, el asistente correspondiente esta herramienta.

Una vez dispuesto el asistente, se elige la opción de situación de los datos («lista o base de datos de Microsoft Office Excel»), y se solicita la creación de una «Tabla Dinámica». Al hacer click en el botón  , el paso 2 muestra directamente el rango de celdas elegido ($A$1:$E$10173). Por último, finalizamos con el paso 3 solicitando que el informe de tabla dinámica que vamos a confeccionar se ubique en una nueva hoja de cálculo que se incluirá en el Libro Excel abierto. A continuación, como vemos en la siguiente figura, en esta nueva hoja de cálculo aparecen: el asistente de creación de la tabla dinámica para la colocación de campos y datos (en las celdas A1:G16), la lista de campos de la tabla dinámica (desde la que podremos arrastrar los campos —variables—), y por último, la barra de herramientas de la tabla dinámica, con su desplegables.

Para construir la tabla dinámica que nos facilite la agregación por hogares de los microdatos individuales hemos de seleccionar la variable (campo) que determina la agregación a realizar: en este caso, el número del hogar («Ho-

gar»). De este modo, si un hogar tiene dos individuos que perciben rentas, puesto que ambos tendrán el mismo número de hogar, sus datos serán tratados de forma agregada, según la opción que elijamos en «Configuración de Campo». Esta selección la realizamos arrastrando con el cursor el campo hasta el asistente de creación, justamente hasta el espacio vertical destinado a «Colocar aquí los campos de fila» (que ocupa las celdas A4:A16). Automáticamente, veremos que todos los números de hogar consecutivos en la muestra, desde el 1 hasta el 8772, ocupan las filas 4 a 8776 de la Columna A.

Una vez establecida la variable de agregación, procedemos a incluir todas las variables cuyos valores deseamos agregar. Para ello, procedemos de igual forma, desplazando con el puntero , y , pero ahora al área destinada a la «Colocación de Datos». Hay que señalar que al incluir dos variables de datos, al incorporar la primera, «Renta Económica», automáticamente éstos se disponen en columna (B), debiendo situar las siguientes variables, «Hijos» y «Cuota Líquida >=0" sobre esta columna, apareciendo automáticamente para cada número de hogar (columna A) tres filas en las columnas B y C, la primera con la opción de agregación y la segunda con el valor correspondiente a ese estadístico para cada variable. Seguidamente, procedemos a comprobar si la opción de agregación que aparece es la que deseamos. Por defecto, la opción elegida habrá sido «Suma» para ambas variables, por lo que ya disponemos directamente de la suma tanto de la «Renta Económica» como de los «Hijos y para la «Cuota Líquida >=0". Si necesitásemos cambiar esta opción de agregación, deberemos marcar en una celda cualquiera de la columna B. Con el botón derecho del ratón abrimos el menú y elegimos la opción . Haciendo click sobre ésta, pueden seleccionarse los distintos estadísticos utilizables como criterios de agregación [5].

Para disponer la información de cada hogar en una única fila, debemos marcar desde arriba todo el bloque de la columna B (la que recoge las opciones de agregación). Para ello, situamos el puntero en la celda B3 (la que recoge el encabezado «Datos» de la tabla dinámica, y al aparecer el puntero en la forma de flecha vertical se marca mediante el botón izquierdo del ratón. Marcando el bloque, arrastramos con el puntero en forma de cruz hasta el encabezado de la columna C, que recogía el «Total». Automáticamente, la información queda dispuesta como deseamos: cada fila corresponde a un hogar, la columna B recoge la renta económica total de ese hogar, la columna C el número total de hijos menores del hogar, y la columna D la cuota líquida del IRPF de ese hogar. Marcando el bloque con toda la

[5] La herramienta de tablas dinámicas ofrece los siguientes criterios estadísticos para agregar la información correspondiente al campo de referencia: suma, cuenta del número de observaciones, media aritmética, valor máximo, valor mínimo, producto, contar números, desviación estándar, desviación estándar poblacional, varianza y varianza poblacional.

información y copiando los valores en otra hoja de cálculo, ya podemos disponer de una base de datos con la información que necesitamos pero para los 8772 hogares (en lugar de para los 10172 individuos). Hay que señalar que la única limitación que presenta la herramienta de Tabla Dinámica puede venir determinada por el número de filas disponibles cuando empleamos el asistente de Creación de la Tabla Dinámica, puesto que como hemos visto, la información va situándose para cada valor del campo de referencia (hogar) en filas consecutivas. Una forma de superar esta limitación es trocear previamente la base de datos original, en tantas partes como se requieran para no superar el número de filas máximo ofrecido por Excel (65536).

Solución al Ejercicio Propuesto n.º 3

1. En primer lugar, sobre la base de datos de los hogares, procedemos a simular la aplicación del programa de ayudas por hijos que no discrimina por nivel de renta de la familia. Para ello, simplemente obtenemos en una columna el importe a percibir por cada hogar, multiplicando la ayuda unitaria de 500,00 euros por el número de hijos del hogar [6]. La suma de la columna G (Ayuda Familiar Alt. 1), 3.391.500,00 euros es el coste presupuestario de aplicar el programa en esta alternativa. Si dividimos este importe entre 500,00 euros, obtenemos el número de hijos recogidos en los hogares de la muestra, 6.783.

2. Puesto que nuestro objetivo en la evaluación propuesta es medir el impacto que esta medida tiene sobre el bienestar social, teniendo en cuenta su efecto redistributivo, tenemos que tener en cuenta en la comparación a realizar cuál es el efecto que el coste presupuestario de la medida tiene sobre la renta media de los hogares de esta sociedad. Si procediésemos a comparar directamente el valor de la FBS antes de aplicar las transferencias que articulan el programa de ayuda familiar con el resultante tras su aplicación, estaríamos considerando un «tamaño del pastel» más grande, pues la sociedad sería más rica, exactamente dispondría de una renta de 3.391.500,00 euros más. Sin embargo esto no es cierto, pues el coste presupuestario de la medida debería ser financiado a través del sistema tributario. Por supuesto, el patrón de distribución de la carga de ese incremento de los impuestos también ha de tener su influencia en el bienestar social, pero no parece demasiado conveniente, al menos en una primera evaluación, mezclar los efectos redistributivos del gasto y del ingreso.

[6] Los casos en los que el número de hijos incorpora una fracción decimal corresponden a hogares en los que todos o alguno de los hijos ha sido incorporado a un nuevo hogar por uno sólo de los padres. Es evidente, que la información disponible impide conocer la diferencia entre una aportación de dos hijos a un nuevo hogar y el caso de un hogar con un solo hijo. No obstante, a efectos del coste presupuestario, como sucede en el IRPF, consideramos la aplicación en función de la distribución de la carga por parte de los padres.

Ante esta situación, ¿cómo podemos actuar para poder llevar a cabo un análisis consistente? La opción habitualmente empleada en los análisis redistributivos consiste en neutralizar el efecto que la medida analizada tiene sobre la renta medida, pero también de forma neutral respecto a su incidencia en la distribución de la renta. Como se ha visto en el capítulo dedicado a la medición de la desigualdad, transformar todas las rentas de una distribución mediante la aplicación de un valor constante, como puede ser el tipo de un impuesto proporcional o de un subsidio proporcional, no altera la desigualdad de dicha distribución, medida a través de cualquier índice de desigualdad relativo.

Por tanto, procedemos a calcular el porcentaje de incremento de la renta que supondría para la totalidad de los hogares —tanto aquellos que tienen hijos como aquellos que no los tienen— la distribución del coste presupuestario de la ayuda. Con ello, estamos obteniendo el incremento de renta *per cápita* que supone la medida, para poder comparar el cambio distributivo aislando el efecto que una mayor renta para los hogares beneficiados tendría sobre el valor de *W*. Dividiendo los 3.391.500,00 euros a pagar entre el total de la renta económica de los hogares, que asciende según la suma de la columna C a 151.591.349,27 euros, obtenemos un incremento de la renta del 2,237265%. Ahora procedemos a aplicar este subsidio proporcional del 2,237265% a la renta antes de impuestos de todos los hogares, lo que nos permite contar con una distribución que tiene la misma desigualdad que la real, previa a la aplicación del programa de ayuda familiar, pero con una renta media igual a la que resultaría tras su aplicación, 17667,90 euros, en lugar de la renta media inicial de 17281,28 euros.

3. Una vez obtenida la distribución «instrumental», ya podemos realizar la comparación con la distribución resultante de aplicar el programa de ayuda familiar en su alternativa 1. Esta distribución resulta de sumar para cada hogar, su renta económica y el importe de la ayuda percibida (las columnas C e I de la base de microdatos). Para ambas distribuciones obtenemos el valor correspondiente a la utilidad de la renta, aplicando a cada renta la función de utilidad propuesta, $U = x^{\frac{1}{4}}$. La suma de utilidades para la distribución instrumental (equivalente en desigualdad a la previa a la introducción de la ayuda familiar) asciende a 1081362,52, mientras que la correspondiente a la distribución resultante tras el programa de gasto en su alternativa 1 asciende a 1082219,48. Empleando la especificación de la FBS en términos de utilidad media propuesta en la expresión [7] obtenemos una mejora del bienestar social de 0,097693, al aumentar *W* de 123,274341 a 123,372034 (un 0,079%).

Puesto que ambas distribuciones de renta tienen la misma media, este resultado se debe estrictamente a la reducción de la desigualdad que ha introducido la ayuda por hijos. De hecho, si obtenemos los índices de Gini

de ambas distribuciones, comprobamos esta reducción: de un Gini de 0,403467 se pasa a un Gini de 0,401073. Por tanto, el programa genera un efecto redistributivo de 0,002394 medido en términos del índice de Reynolds-Smolensky.

4. En segundo lugar, realizamos la evaluación para la alternativa 2 del programa. Puesto que restringimos la aplicación de la ayuda familiar por hijos menores a aquellos hogares con una renta no superior a 20.000 euros, ahora el coste del programa se ve reducido a 1.929.000,00 euros. Con este gasto, ahora el tipo de subsidio proporcional a aplicar a todos los hogares para obtener la distribución instrumental pasa a ser el 1,2725% de la renta antes de impuestos. Ahora, por tanto, la renta media de las dos distribuciones comparadas se eleva a 17501,18 euros. Obteniendo las utilidades de la renta de cada hogar para ambas distribuciones, calculamos las respectivas sumas de utilidades de la renta antes y después de aplicar el programa de ayuda en su alternativa 2: 1076248,27 y 1078192,82. Por consiguiente, para la FBS utilizada obtenemos una mejora del bienestar social de 0,221677. (al aumentar un 0,181% la utilidad media, desde 122,691322 hasta 122,9130).

De acuerdo con el proceso seguido, este aumento en el bienestar social que consigue la aplicación del programa de ayuda familiar en su segunda modalidad es consecuencia de la reducción en la desigualdad que introduce el sistema de transferencias contemplado. En este caso, el índice de Gini se reduce desde su valor inicial del 0,403467 (no olvidemos que la transformación realizada para obtener la distribución instrumental no afecta a la desigualdad de la distribución original) hasta un valor de 0,397652. El efecto redistributivo es ahora sustancialmente mayor, 0,005815, más del doble que el alcanzado por la alternativa 1, aunque conseguida con un nivel de gasto notablemente menor.

5. Para concluir la evaluación planteada, nos queda comparar la ordenación entre las dos alternativas consideradas. Como hemos visto, amabas resultarían aprobadas por una FBS individualista y con aversión a la desigualdad como la empleada, pues su aplicación proporciona ganancias de bienestar social ($\Delta W > 0$). Sin embargo, la forma en la que se distribuyen los hijos en los hogares a lo largo de la distribución de renta (la media de hijos aumenta conforme aumentamos la decila de renta) amortigua el efecto redistributivo de este gasto en la alternativa 1, al concentrarse un número importante de hijos en los hogares con rentas elevadas. Por tal motivo, la restricción de aplicación de la alternativa 2, a los hogares con rentas no superiores a 20.000 euros introduce una fuerte progresividad en la ayuda, que compensa la reducción del gasto presupuestado. Esta hace que la reducción de la desigualdad en la segunda

opción sea mucho más intensa. Consecuentemente, si medimos la mejora del bienestar en términos relativos —como cambio porcentual respecto del nivel previo a la introducción del programa, puesto que la influencia sobre la renta media al tener diferentes coste presupuestario es distinta—, como vimos, la ganancia de bienestar social es mayor para la alternativa 2, 0,1807% frente al 0,0792%. En definitiva, el análisis recomienda la aplicación de la alternativa 2 del programa de ayuda familiar por hijos.

El criterio de dominancia de Lorenz

Como hemos visto en el capítulo 5, las curvas de Lorenz permiten analizar gráficamente, de forma muy intuitiva, la desigualdad de una distribución de la renta, así como realizar comparaciones de desigualdad entre distribuciones. Este análisis puede extenderse a la comparación de curvas de Lorenz correspondientes a las distribuciones de la renta antes y después de aplicar un determinado impuesto o un programa de transferencias de gasto específico. Este método también nos permite comparar en el tiempo la evolución de la desigualdad de la renta o sus diferencias territoriales, por ejemplo, entre países o regiones.

Además de esta indiscutible utilidad, Atkinson (1970) encontró una interesantísima potencialidad de carácter normativo a la comparación entre curvas de Lorenz. Este resultado se recoge en el conocido Teorema de Atkinson, según el cual si la curva de Lorenz $L_X(p)$ correspondiente a la distribución $F(X)$ se sitúa por encima de la curva de Lorenz $L_Y(p)$ correspondiente a la distribución $F(Y)$, a lo largo de todos los percentiles acumulados de población (p), podemos decir que para cualquier FBS individualista y con aversión a la desigualdad (como la expuesta en [7]) el nivel de bienestar social alcanzado con la distribución $F(X)$ será siempre superior al que se obtendría con $F(Y)$.

Analíticamente, esta condición geométrica exigida a las curvas de Lorenz comparadas exige que $L_X(p) \geq L_Y(p)$. Por la propia definición de la curva de Lorenz, resulta trivial que para p=0 y para p=1, $L_X(p) = L_Y(p)$. La verificación de esta relación recibe la denominación de *dominancia de Lorenz* o *dominancia en el sentido de Lorenz* (ver figura 10). Puesto que como hemos visto, la FBS que asegura el respeto al principio de aversión a la desigualdad debe ser cóncava, este criterio de dominancia coincide con el criterio estadístico de *dominancia estocástica de segundo orden* (ver Rodríguez y Salas, 2003).

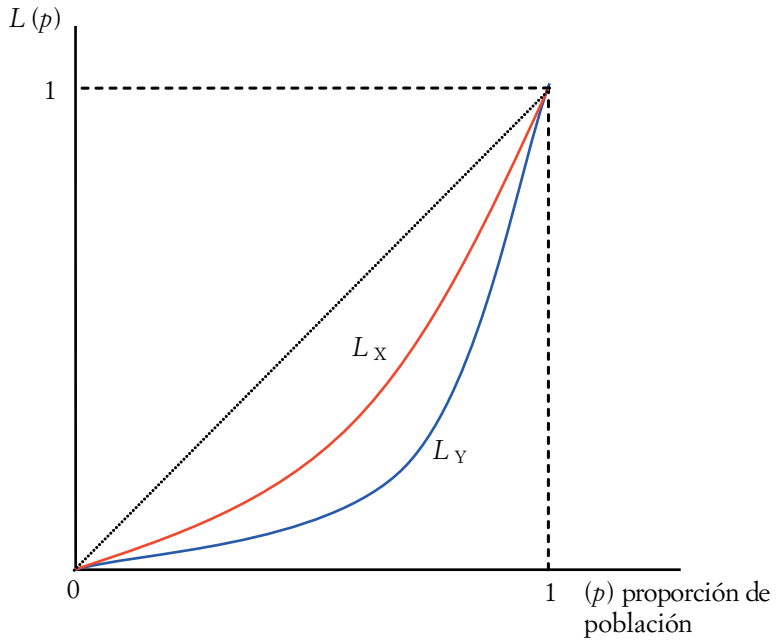


FIGURA 10

La aplicación de este resultado exige, en principio, una condición que para cierto tipo de evaluaciones podemos considerar como bastante restrictiva: las dos distribuciones cuyas curvas de Lorenz son sometidas a comparación deben poseer idéntica media. Como señala Lambert (2001:44), la curva de Lorenz muestra cómo se divide el pastel, pero no revela nada acerca de su tamaño ni del número de ciudadanos entre los que se reparte.No olvidemos que se trata de una función definida en términos de fracción de una variable —generalmente la renta— y con un soporte que es la fracción de la población que detenta las respectivas fracciones de dicha variable. Sin embargo, como ya razonamos al inicio de esta sección, el tamaño del pastel es relevante a la hora de juzgar el bienestar del que disfruta la sociedad. Por tanto, el valor medio de la renta resulta fundamental. Resulta evidente, que si comparamos formas alternativas de distribuir el mismo pastel entre el mismo número de comensales —es decir, cuando esas distribuciones comparadas tienen la misma media— el problema desaparece.

En el análisis de políticas públicas, una forma habitual y bastante ocurrente de evitar este problema consiste en realizar la comparación no entre las distribuciones de la renta anterior y posterior a la aplicación de esa intervención, sino entre la distribución de la renta resultante de su aplicación y la distribución hipotética que se alcanzaría si la medida hubiese sido aplicada de forma estrictamente proporcional. Como hemos expuesto en la resolución del ejerci-

cio n.º 3, al multiplicar todas las observaciones de una distribución por un valor constante (un escalar), la desigualdad relativa de la distribución no se ve afectada. De este modo, por ejemplo, si comparamos la introducción de un determinado impuesto progresivo, con un tipo medio efectivo t_{PG}, la renta media antes de impuestos de los contribuyentes, μ_X, pasaría a ser $\mu_Y = \mu_X \cdot (1 - t_{PG})$. Aunque de acuerdo con el carácter progresivo de ese impuesto cada contribuyente soportará un tipo medio distinto, creciente con el nivel de renta, un impuesto proporcional que generase la misma recaudación tendría un tipo medio efectivo t_P idéntico a t_{PG}, aunque como hemos visto, su capacidad redistributiva sería nula. De este modo, se consigue que ambas distribuciones comparadas posean igual media, centrándose el análisis de bienestar social en la potencialidad de la estructura progresiva para reducir la desigualdad. En otros casos, como la posible reforma del tratamiento de la unidad familiar del impuesto o de las cargas por dependientes, los análisis suelen realizarse bajo el supuesto de ausencia de coste recaudatorio, lo que permite evitar también esta limitación impuesta por la exigencia de igualdad de medias.

No obstante, el contenido del Teorema de Atkinson es también de aplicación a la comparación de distribuciones con diferente media, siempre y cuando coincida que la distribución dominante en sentido de Lorenz posea también una media mayor. Se trata un corolario del Teorema bastante intuitivo si nos atenemos a los argumentos expuestos. Si partimos de que la concavidad de la FBS asegura ganancias de bienestar en la medida que la desigualdad se reduce, si además añadimos que la distribución con menor desigualdad reparte porciones del pastel de mayor tamaño es inmediato que el valor de la utilidad media de la sociedad será superior. En términos analíticos, la comparación en términos de Lorenz de dos distribuciones como la presentada, $L_X(p) \geq L_Y(p)$, con medias μ_X y μ_Y tal que $\mu_X > \mu_Y$, puede resolverse de la siguiente forma: la distribución de las rentas $F(X)$ es preferida a distribución (μ_X/μ_Y) de veces las rentas en $F(Y)$.

Para ilustrar la aplicación del criterio de dominancia de Lorenz y el teorema de Atkinson (1970) proponemos el siguiente ejercicio que resolvemos.

EJERCICIO PROPUESTO N.º 4 (SOLUCIONADO)

En la sociedad cuya distribución de la renta se ofrece (Fichero Muestra IRPF Individuos - Hogares con Hijos), se plantea sustituir el actual impuesto progresivo sobre la renta articulado a través de una tarifa con varios tipos marginales crecientes y diversas reducciones y deducciones de la cuota por un impuesto lineal con un mínimo exento de 6.000,00 euros y un único tipo marginal, cuya cuantía debe asegurar el mismo nivel de recaudación que el impuesto precedente. El nuevo impuesto está previsto que tenga una concepción estrictamente individual por lo que no se contempla ningún tipo de deducción familiar. Se pide evaluar en términos de bienestar social esta reforma tributaria.

Para realizar esta evaluación empleamos el fichero con datos individuales correspondientes a la aplicación del IRPF. Puesto que el nuevo impuesto considerado no incorpora ningún tipo de reducciones o deducciones, tan sólo necesitamos disponer de las variables correspondientes a la renta antes de impuestos (Renta Económica) y a la cuota generada por el anterior impuesto (Cuota Líquida).

Solución al Ejercicio Propuesto n.º 4

1. En primer lugar, procedemos a simular la aplicación de la nueva estructura tributaria que sustituirá al IRPF aplicado hasta la fecha. Como se ha señalado en el enunciado este nuevo impuesto tiene la estructura propia de un *Flat Tax*, incorporando un mínimo exento de 10.000,00 euros por declarante. Para calcular el tipo marginal aplicable a partir del umbral de este mínimo exento, emplearemos la hoja de cálculo Excel introduciendo la siguiente expresión determinante de la cuota líquida del nuevo impuesto: $CL = (X - M) \cdot t_{mg}$, donde X representa la renta económica gravable, M el mínimo exento de 10.000 euros, y t_{mg} el tipo marginal buscado.

Para buscar el tipo marginal del nuevo impuesto, utilizamos como referencia del nivel recaudatorio del IRPF vigente su tipo medio efectivo. El cociente entre la suma de cuotas líquidas de este impuesto y la suma de las rentas económicas nos ofrece este valor, 11,918%. Fijando el mínimo exento en 10.000 y estableciendo la condición de que para cualquier renta económica por debajo de estos 10.000 euros, la cuota líquida será igual a cero, aproximamos distintos valores para t_{mg}, hasta encontrar $t_{mg} = 0{,}2795$, lo que proporciona un tipo medio efectivo del 11,918%, que asegura la neutralidad recaudatoria del medida.

2. Una vez definida la estructura del nuevo impuesto ($CL = (X - 10.000) \cdot 0{,}2795$), obtenemos las cuotas líquidas correspondientes a los 10172 declarantes que componen nuestra muestra. Como podemos comprobar, la suma total de cuotas líquidas del nuevo impuesto simulado es prácticamente igual a la alcanzada por el IRPF que se pretende sustituir, aproximadamente 18.067.000 euros.

Para responder a la cuestión planteada es necesario comparar las curvas de Lorenz correspondientes a las dos distribuciones de la renta netas del impuesto personal. De acuerdo con el contenido de la reforma formulada, como hemos visto ambas distribuciones de renta presentan el mismo valor en su media, 13.126,6 euros. Por tanto, la utilización del criterio de dominancia de Lorenz para evaluar la deseabilidad de la reforma en términos de bienestar social es, en principio, pertinente.

La utilización de Excel para el análisis plantea la posibilidad de obtener directamente las curvas de Lorenz para una división poblacional equivalente a la proporción de cada observación, $1/N$. Sin embargo, esta opción no es demasiado recomendable pues resulta complicado identificar la posible intersección de las curvas comparadas. Normalmente, se suele proceder a analizar la dominancia agrupando las observaciones de la renta de cada variable por centilas, operando para cada una de ellas con el valor medio de la renta en la centila. Para desarrollar este método en primer lugar procedemos a ordenar la distribución de cada variable (en nuestro caso, la Renta Neta del IRPF y la Renta Neta del Flat Tax) en sentido ascendente. A continuación, como ya hemos visto en la presentación de la curva de Lorenz (capítulo 5), procedemos a identificar en una columna la centila en la que se ubica cada observación mediante la expresión "=ENTERO(A4/(10172/100))+1", donde A4 representa el dígito de posición de esa celda de acuerdo con la ordenación creciente de la variable, siendo 10172 el número de observaciones (celdas) que contiene la distribución de la variable correspondiente. No hay que olvidar corregir en la última observación el valor devuelto por esta expresión, que será por construcción 101, debiendo ser 100 (el de la última centila).

Una vez obtenida la centila a la que pertenece cada observación de la renta neta que evaluamos, empleamos nuevamente la herramienta de Tabla Dinámica para obtener el valor medio de la renta en cada centila. Ahora, la variable de referencia será precisamente «centila», eligiendo la opción estadística de «promedio». Una vez obtenidos estos valores medios de las centilas de la 1 a la 100 para las dos distribuciones comparadas, procedemos a calcular las fracciones acumuladas de renta en cada una de ellas, para así obtener el valor correspondiente de la curva de Lorenz, L_X y L_Y en cada centila. Puesto que (p) es común a ambas distribuciones de Lorenz, tan solo hemos de comprobar si los valores de L correspondientes a una de las curvas son superiores a los de la otra, lo que supone la inexistencia de intersecciones.

En el ejercicio realizado, vemos que hasta la centila 5, la curva de Lorenz correspondiente a la distribución de la renta neta de la aplicación del IRPF domina (va por encima) de la curva de Lorenz de la distribución de la renta neta del impuesto lineal propuesto, aunque prácticamente coinciden (las diferencias se resuelven a nivel del séptimo decimal). A partir de la sexta centila, ambas curvas se cruzan, pasando a dominar la curva de Lorenz asociada a la aplicación del nuevo impuesto, hasta la centila 85, en la que de nuevo la curva de Lorenz resultante de la aplicación del IRPF vigente se sitúa por encima.

El resultado obtenido en la comparación, con intersecciones en las curvas de Lorenz, no nos permite obtener una preferencia unáni-

	A	B	C	D	E	F	G	H	I
1	**Centila**	**L(RN IRPF)**	**L(RI Flat Tax)**	**Diferencia**		**Centila**	**L(RN IRPF)**	**L(RI Flat Tax)**	**Diferencia**
2	1	0,00012311	0,00012308	0,0000000		26	0,09849869	0,098398681	-0,0009869
3	2	0,00082292	0,00082915	0,0000000		27	0,10441686	0,105427822	-0,0010108
4	3	0,0021371	0,00213709 2	0,0000000		28	0,1104337	0,111580531	-0,0011468
5	4	0,00392004	0,00392002 9	0,0000000		29	0,11649686	0,11777376	-0,0012759
6	5	0,00613447	0,00613444 3	0,0000000		30	0,1227270	0,124145186	-0,0014174
7	6	0,00869157	0,00869244 1	-0,0000009		31	0,12906341	0,130642539	-0,0015791
8	7	0,01157114	0,01157707 3	-0,0000059		32	0,13550384	0,137269386	-0,0017655
9	8	0,01468835	0,01470751 9	-0,0000192		33	0,1419816	0,143939628	-0,0019580
10	9	0,01805259	0,01809764 9	-0,0000451		34	0,14862426	0,150780508	-0,0021662
11	10	0,02163605	0,02170364 3	-0,0000676		35	0,15537263	0,157731546	-0,0023809
12	11	0,02533701	0,02542381 1	-0,0000868		36	0,16214966	0,164741439	-0,0025918
13	12	0,02923786	0,02934297 5	-0,0001051		37	0,16908751	0,171937442	-0,0028469
14	13	0,03329011	0,03342117 2	-0,0001311		38	0,17612438	0,179244387	-0,0031200
15	14	0,03748946	0,03764719 1	-0,0001577		39	0,18325325	0,186647385	-0,0033941
16	15	0,04177325	0,04196379	-0,0001905		40	0,1904104	0,194082537	-0,0036721
17	16	0,04624554	0,04647267 6	-0,0002274		41	0,19773101	0,201701 06	-0,0039701
18	17	0,05088698	0,05112659 5	-0,0002596		42	0,20514792	0,209416529	-0,0042689
19	18	0,05558788	0,05587979 8	-0,0002921		43	0,21269397	0,217155232	-0,0045613
20	19	0,06049309	0,06082123 3	-0,0003281		44	0,22024317	0,225066713	-0,0048225
21	20	0,0655433	0,06592020 5	-0,0003759		45	0,22799393	0,23306949	-0,0050656
22	21	0,07073713	0,07117229 7	-0,0004352		46	0,23605651	0,241147016	-0,0052905
23	22	0,07601978	0,07651634 3	-0,0004996		47	0,2437494	0,249249028	-0,0054986
24	23	0,08148028	0,08204725 7	-0,0005670		48	0,25183937	0,257524981	-0,0056856
25	24	0,08705827	0,08771173 3	-0,0006535		49	0,26005803	0,265897968	-0,0058339
26	25	0,09269686	0,09345724 9	-0,0007814		50	0,26830456	0,274292609	-0,0059881
27									
28	**Centila**	**L(RN IRPF)**	**L(RI Flat Tax)**	**Diferencia**		**Centila**	**L(RN IRPF)**	**L(RI Flat Tax)**	**Diferencia**
29	51	0,27675293	0,282883346	-0,0061304		76	0,53643526	0,540266516	-0,0038303
30	52	0,28533356	0,291579692	-0,0062461		77	0,54932823	0,552850738	-0,0035245
31	53	0,2940357	0,300387215	-0,0063515		78	0,56247283	0,565702638	-0,0032239
32	54	0,30276913	0,309212743	-0,0064436		79	0,57575715	0,578660075	-0,0029109
33	55	0,31171091	0,318239704	-0,0065269		80	0,5894689	0,592026076	-0,0025508
34	56	0,32078802	0,327381241	-0,0065932		81	0,60345512	0,606661437	-0,0023053
35	57	0,32999804	0,336837215	-0,0066403		82	0,61773901	0,619568261	-0,0018203
36	58	0,33925607	0,346927031	-0,0067110		83	0,63220636	0,63362226	-0,0014159
37	59	0,3487656	0,356431404	-0,0067758		84	0,64716243	0,648110392	-0,0009480
38	60	0,35841994	0,366076056	-0,0068559		85	0,66245344	0,662916656	-0,0004652
39	61	0,36813092	0,374757434	-0,0068305		86	0,67793671	0,677924616	0,0000121
40	62	0,37810587	0,384876102	-0,0067702		87	0,69393727	0,693475905	0,0004814
41	63	0,38823745	0,394740905	-0,0069035		88	0,71034957	0,709410548	0,0009399
42	64	0,39853858	0,404861496	-0,0070229		89	0,72727433	0,725766125	0,0015082
43	65	0,40891989	0,415224606	-0,0073046		90	0,74452348	0,742306786	0,0021367
44	66	0,41956853	0,425752246	-0,0078537		91	0,76249397	0,759769993	0,0027280
45	67	0,43043814	0,436438619	-0,0080026		92	0,78111459	0,777837099	0,0032775
46	68	0,44137116	0,447177316	-0,0080852		93	0,80032709	0,796536089	0,0037910
47	69	0,45261181	0,458200574	-0,0058838		94	0,82059821	0,816278403	0,0043198
48	70	0,46402279	0,469391556	-0,0053606		95	0,84195084	0,837033894	0,0049169
49	71	0,47561008	0,480761434	-0,0051505		96	0,8645529 2	0,8591395 3	0,0054134
50	72	0,48727482	0,492200846	-0,0049280		97	0,88873058	0,882944806	0,0057858
51	73	0,49927924	0,503944083	-0,0046717		98	0,91609654	0,910130215	0,0059663
52	74	0,51149047	0,51588002	-0,0043895		99	0,94819351	0,942867926	0,0053256
53	75	0,52360121	0,52790726	-0,0041050		100	1	1	0,0000000

me por una de las alternativas consideradas. En este tipo de evaluación, en los que ambas distribuciones presentan la misma media, la única forma de resolver la ambigüedad que ofrece el análisis de la dominancia de Lorenz es proceder al estudio empírico a través de especificaciones concretas de las FBS con diferentes grados de aversión a la desigualdad (a través de distintos grados de concavidad en las funciones de utilidad).

El criterio de dominancia de Lorenz generalizada

Una alternativa propuesta por Shorrocks (1983) para tratar de extender el criterio de dominancia de Lorenz a aquellas comparaciones de distribuciones que presentan media distinta parte de la siguiente transformación de la curva de Lorenz:

$$GL_x(p) = \mu_x \cdot L_x(p) \qquad [11]$$

Esta transformación, denominada *Curva de Lorenz Generalizada* consiste en multiplicar cada valor (punto) de la curva de Lorenz por la media de la distribución. Shorrocks (1983) demuestra que la dominancia estocástica de segundo orden puede enunciarse en términos de *dominancia de Lorenz generalizada*, en la medida que la dominancia de Lorenz es simplemente un caso particular en la que el valor medio de la variable es 1. Por consiguiente, disponemos así de un criterio que proporciona recomendaciones de bienestar social robustas ante la comparación de distribuciones con distinta media y, lo que es también muy interesante, ante situaciones en las que las curvas de Lorenz se cruzan.

Gráficamente, la dominancia de Lorenz generalizada se produce en los términos recogidos en la figura 11.

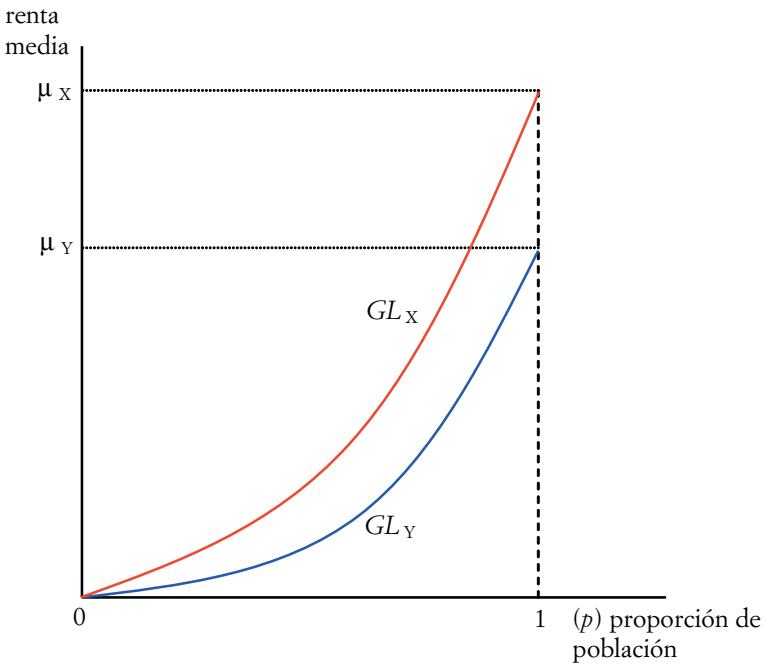


FIGURA 11

De acuerdo con estas posibilidades que nos brinda la comparación de curvas de Lorenz generalizadas, ¿podríamos resolver la ambigüedad a la que nos ha llevado la dominancia de Lorenz en la evaluación impositiva del ejercicio 4? La respuesta es no. En la medida que ambas distribuciones de renta neta poseían la misma media (μ_X / μ_Y), generalizar las curvas de Lorenz no resuelve nada, pues tan sólo supondría un cambio de escala que mantendría las posiciones relativas de ambas inalteradas. Como se analiza en detalle en Lambert (2001), la utilidad de la generalización de las curvas de Lorenz ante intersecciones de las mismas posee utilidad cuando tenemos distribuciones tales que $\mu_X > \mu_Y$ y a su vez se produce que $L_X(p) \leq L_Y(p)$, y la generalización de Shorrocks muestra una dominancia de la distribución con mayor media, tal que $GL_X(p) \geq GL_Y(p)$. En este sentido, debemos destacar, como se desprende de la figura 10, que por la construcción de las curvas de Lorenz generalizadas, nunca la distribución con menor media podrá dominar a la distribución con mayor media, pues para el acumulado poblacional total (p=1), siempre la curva de Lorenz generalizada de la distribución con media superior ira por encima. Otra alternativa que puede ser resuelta mediante la generalización de las curvas de Lorenz se da en aquellos casos en las que las curvas de Lorenz (con distinta media) se cruzan, pero al ser transformadas en curvas de Lorenz generalizadas esta intersección desaparece.

No obstante, en algunos de los casos en los que simple aplicación de la generalización de Shorrocks no permite resolver la ambigüedad en la comparación de bienestar social, la literatura provee algunas respuestas que, aún restringiendo la generalidad de la preferencia unánime a un subconjunto de FBS con aversión a la desigualdad (en función de las varianzas y de la aceptación del *Principio de Transferencias Decrecientes*), ofrecen algunas recomendaciones de elección entre alternativas [7].

El criterio de dominancia de Lorenz generalizada resulta especialmente atractivo a la hora de comparar el nivel de bienestar social entre países o regiones. En estos casos, es evidente que estaremos antes niveles de renta *per cápita* distintos lo que nos sitúa ante distribuciones de media diferente. La combinación de eficiencia —nivel de la renta media— y equidad —grado de desigualdad en la distribución de la renta— puede resolverse mediante la comparación entre curvas de Lorenz generalizadas, respetando la posición ética que admite la aversión hacia la desigualdad. Sin embargo, su aplicación al análisis de políticas públicas resulta, en buena medida, insatisfactorio, pues como hemos visto, las diferencias de medias

[7] Ver la posibilidad de resolución de cruces únicos y de múltiples cruces de las curvas de Lorenz generalizadas en Lambert (2001).

generadas por el distinto coste presupuestario o el impacto recaudatorio de la medida poseen en sí mismas un valor que influye en el bienestar que no puede ser obviado. Métodologías basadas en las igualación del coste recaudatorio o en la neutralización del impacto redistributivo del diferencial de renta son más satisfactorias para este tipo de análisis, como se argumenta en Onrubia, Rodado, Díaz de Sarralde y Pérez (2006).

Evaluaciones mediante funciones de bienestar social abreviadas (no individualistas)

Una alternativa, no exenta de crítica, consiste en la realización de comparaciones de bienestar social a través de funciones de bienestar social abreviadas del tipo,

$$W = \phi(\mu_x, I_x) \qquad [12]$$

donde I_X representa un índice de desigualdad normativo (pudiendo ser el índice de Gini, G_X), y donde f representa cualquier transformación monótona creciente. En concreto, una especificación que goza de interesantes propiedades ético-normativas (ver Dutta y Esteban, 1992) es,

$$W(x) = \phi(\mu_x(1 - I_x)) \qquad [13]$$

donde el índice de desigualdad relativo puede venir representado por el *Indice de Atkinson*,

$$I_x^A(\alpha) = 1 - \left[\sum_{i=1}^{N} \frac{1}{N}\left(\frac{x_i}{\mu_x}\right)^{1-\alpha}\right]^{\frac{1}{1-\alpha}}; \forall \alpha > 0 \qquad [14]$$

en el que α representa un parámetro que recoge el grado de aversión a la desigualdad que muestra la FBS subyacente, de forma que cuando α aumenta su valor, la sensibilidad del índice ante las transferencias en la parte baja de la distribución es mayor, en detrimento de las que tienen lugar en la parte alta.

Este tipo de FBS tienen su origen en la necesidad apuntada de ordenar distribuciones de la renta para las cuales el criterio de dominancia de Lorenz generalizada no ofrece preferencias unánimes. La idea se basa en los trabajos pioneros de Atkinson (1970), Breit (1974) y Kondor (1975) y consiste, esencialmente, en la búsqueda de un criterio de evaluación social

a partir de la relación de intercambio existente entre el principio de eficiencia —recogido a través de la variación en la renta media— y el grado de desigualdad de la distribución de la renta, recogido a través de un índice de desigualdad. Hay que destacar, no obstante, que se trata de FBS no individualistas, aunque como algunos autores han tratado de demostrar, en algunos casos sus propiedades son compatibles con los resultados obtenidos bajo una construcción individualista.

La habitualidad en la utilización del índice de Gini, ha llevado a la utilización frecuente de este índice en este tipo de FBS abreviadas a pesar de la existencia de algunas críticas respecto de su contenido normativo. En este sentido, una de las especificaciones de estas funciones de evaluación social más empleada es la siguiente:

$$W(x)=\mu_x\cdot(1-k\cdot G_x);\ 0<k\leq 1 \qquad [15]$$

Sheshinski (1972) muestra que [12] es consistente con los requisitos axiomáticos de individualismo y simetría, para unas preferencias individuales representadas por una función de utilidad tal que $\partial U/\partial x>0$ y $\partial^2 U/\partial x^2<0$, y donde el parámetro *k* estaría recogiendo la sensibilidad pro igualdad del evaluador social.

Una interesante interpretación de este parámetro es realizada por Lambert, en términos de *índice de envidia/altruismo* de los ciudadanos. La elasticidad de intercambio entre la renta media (eficiencia) y el grado de desigualdad (equidad) con la que la renta (u otra variable económica) se distribuye dependerá, obviamente, de este parámetro, lo que introduce un indiscutible juicio de valor para su aplicación. No obstante, dando valores a lo largo de todo su rango a este parámetro podemos ofrecer un panorama suficientemente amplio de posiciones éticas respecto del intercambio «eficiencia-equidad» que subyace a cada distribución. Para una especificación como la mostrada en [12], esta elasticidad se define como:

$$\eta^{\mu_x G_x}=\frac{G_x}{\mu_x}\cdot\left.\frac{\Delta\mu_x}{\Delta G_x}\right|_W=\frac{k\cdot G_x}{1-k\cdot G_x} \qquad [16]$$

Como práctica para aplicar este tipo de funciones de bienestar social abreviadas, proponemos el siguiente ejercicio en el que se trata de establecer una ordenación de niveles de bienestar social entre una serie de países de los que disponemos de su renta per cápita y de una medida de desigualdad en la distribución de su renta como es el índice de Gini.

EJERCICIO PROPUESTO N.º 5

Establecer un ranking de países de acuerdo con el concepto de bienestar social recogido en la especificación de la FBS propuesto en la expresión [12]. Para evaluar la posición ética del decisor social emplee distintos valores de parámetro de «sensiblidad pro igualdad». La información disponible para elaborar se recoge en la siguiente tabla.

TABLA EJERCICIO N.º 5

Países	*PIB per cápita (USD) 2006*	*Indice de Gini (Renta)*
Albania	2.899	28,2
Bosnia-Herzegovina	2.533	26,2
Colombia	2.888	58,6
Túnez	2.982	39,8
Marruecos	1.886	39,5
Gambia	325	50,2
Honduras	1.213	53,8
Egipto	1.489	34,4
Paraguay	1.483	57,8
Brasil	5.717	54,0
Ecuador	2.987	43,7
Perú	3.374	54,6

Fuente: Fondo Monetario Internacional.

BIBLIOGRAFÍA

Atkinson, A. B. (1970). «On the measurement of inequality», *Journal of Economic Theory*, 2: 244-263.

Breit, W. (1974). «Income redistribution and efficiency norms», en H. M. Hochman y G. E. Peterson, *Redistribution Trough Public Choice*. New York: Columbia University Press.

Dupuit, J. (1844). «On the measurement of the utility of public works», en K. J. Arrow y T. Scitovsky (eds.), *Readings in Welfare Economics*, London: Allen and Unwin, pp. 255-283.

Dutta, B. y J. M. Esteban (1992). «Social welfare and equality», *Social Choice and Welfare*, 50: 49-68.

Harberger, A. C. (1971). «Three basic postulates for applied Welfare Economics», *Journal of Economic Literature*, IX: 785-797.

Hausman, J. A. (1981). «Exact consumer's surplus and deadweight loss», *The American Economic Review*, 71: 622-676.

Hicks, J. (1939). *Value and Capital*. Oxford: Clarendon Press.

Kay, J. A. y M. A. King (1984). *The British Tax System*, 3.ª ed., Oxford: Oxford University Press.

Kondor, Y. (1975). «Value judgements implied by the use of various measures of income inequality», *Review of Income and Wealth*, 21: 309-321.

Lambert, P. J. (2001). *The Distribution and Redistribution of Income*. Manchester: Manchester University Press. Existe traducción al castellano, *La distribución y redistribución de la renta*. Madrid: Instituto de Estudios Fiscales, 1996.

Onrubia, J., M. C. Rodado, S. Díaz de Sarralde y C. Pérez (2006). «Progresividad y redistribución a través del IRPF español: un análisis de bienestar social para el periodo 1982-1998», *Hacienda Pública Española/Revista de Economía Pública,* 183. Próxima publicación.

Onrubia, J., R. Salas y J. F. Sanz (2005). «Redistribution and Labour Supply», *Journal of Economic Inequality*, 3: 109-124.

Onrubia, J. y J. F. Sanz (2003) (eds.). *Redistribución y Bienestar a través de la Imposición sobre la renta Personal*. Madrid: Instituto de Estudios Fiscales.

Rodríguez, J. G. y R. Salas (2003). «Dominancia estocástica e inferencia estadística en el análisis de las reformas fiscales», en J. Onrubia y J. F. Sanz (eds.), *Redistribución y bienestar a través de la imposición sobre la renta personal*. Madrid: Instituto de Estudios Fiscales, pp. 59-90.

Rosen, H. S. (2005). *Public Finance*. Richard D. Irwin. Existe traducción al castellano de la 5ª edición, *Hacienda Pública*, McGraw-Hill, 2002.

Sheshinski, E. (1972). «Relation between a social welfare function and the Gini index of income inequality», *Journal of Economic Theory*, 4: 98-100.

Shorrocks, A. (1983). «Ranking income distributions», *Economica*, 50:1-17.

Willig, R. (1976). «Consumer's surplus without apology», *The American Economic Review*, 66: 589-597.

CAPÍTULO IX

EFICIENCIA DE LAS POLÍTICAS DE FAMILIA MEDIANTE ANÁLISIS ENVOLVENTE DE DATOS

DANIEL SANTÍN GONZÁLEZ

9.1. INTRODUCCIÓN

Las políticas públicas de familia, la lucha contra la pobreza o los programas favorecedores del bienestar social a través de la educación o la sanidad han aumentado su dimensión considerablemente en los países desarrollados. En su mayoría estas políticas son llevadas a cabo por diferentes unidades gestoras; como centros de asistencia a la familia, distritos, ayuntamientos o equipos de trabajo que se enfrentan a la tarea de alcanzar los mejores resultados posibles dado un presupuesto. En este contexto cabe preguntarse cuál es el mejor procedimiento de gestionar los recursos para maximizar los logros planteados dados a las unidades gestoras, o lo que es lo mismo, dados unos objetivos cuál es «el mecanismo» que permite obtener esos resultados al menor coste. Cuando una unidad productiva logra este «resultado» decimos que es eficiente.

Tradicionalmente el sector público ha evaluado su actuación utilizando el concepto de eficacia. Se puede definir como eficacia el cumplimiento de un objetivo. Así, si el sector público tiene como objetivo informar a mil familias y se cumple con el mismo a lo largo del ejercicio diremos que la actuación ha sido eficaz. También es frecuente hablar del grado de eficacia o del grado de cumplimiento de un objetivo, cuando se relaciona el objetivo logrado con el objetivo inicialmente previsto. En el caso anterior si se preveía informar a 1000 familias y únicamente se ha informado a 900 diríamos que la eficacia de la política ha sido del 90%. El grado de eficacia también se utiliza a menudo para analizar la evolución de los resultados a lo largo del tiempo. Sin embargo la definición de eficacia no tiene en cuenta los medios humanos y materiales utilizados para lograr los objetivos por lo que el concepto de eficiencia va más allá al relacionar factores productivos o recursos (inputs) con objetivos de resultados (outputs).

La evaluación a través del concepto de eficiencia puede aplicarse tanto a empresas privadas como a unidades de gestión pública. En nuestro caso estamos interesados en la eficiencia de las políticas públicas por lo que aunque en ocasiones la producción de estas políticas puede ser privada (Organizaciones No Gubernamentales, Fundaciones, etc.) en la mayoría de casos las unidades gestoras serán de titularidad pública.

La medición de la eficiencia en el sector público es en la mayoría de ocasiones un proceso aún más complejo que en el sector privado. Ello es debido a que el comportamiento microeconómico del sector público se ve afectado por las siguientes características.

1. En primer lugar, los objetivos que el político transmite a la unidad gestora son en numerosas ocasiones múltiples, complejos y en ocasiones difusos y no claramente definidos. En cualquier caso estos objetivos distan de estar relacionados con la maximización del beneficio económico ya que estas políticas están orientadas hacia la búsqueda de la rentabilidad social.

2. En segundo lugar, la actuación pública carece de mercados en competencia perfecta que determinen los precios. ¿Cuánto vale ayudar a una familia a salir de la pobreza?, ¿Cuál es el precio de reducir el tiempo de espera desde que se solicita una ayuda hasta que se concede? Lógicamente, al contrario que en la empresa privada no existe un mercado que asigne un precio a estos resultados de la producción de las unidades encargadas de la gestión de las políticas sociales.

3. En tercer lugar, en la función pública los trabajadores no suelen tener incentivos claros ligados a objetivos y no es fácil la movilidad en los puestos de trabajo. En este contexto puede producirse lo que Leibenstein (1966) denominó como *Ineficiencia X*. Así, dado que los trabajadores no tienen un incentivo claro para maximizar su productividad podrían maximizar su utilizad limitando su esfuerzo y utilizando más presupuesto que el necesario.

A pesar de estas dificultades, resulta clave cuantificar la ineficiencia media y la ineficiencia individual del proceso productivo de las políticas de familia llevado a cabo por distintas unidades gestoras. Esta herramienta de evaluación no debe entenderse en ningún caso como una manera de fiscalizar la actividad de los centros gestores sino como un mecanismo para:

1. Diagnosticar posibles dificultades en la gestión debidas a circunstancias que pueden ser muy diversas y que en ocasiones escapan al control de los productores.

2. Detectar las mejores prácticas productivas. A partir de ellas podemos aprender de su gestión y extrapolarla a aquellas unidades que en un contexto parecido son ineficientes.

3. La medición de la eficiencia ofrece objetivos a las unidades gestoras ineficientes en términos de cuáles serían los resultados a alcanzar (o el recorte presupuestario que debería realizarse) para ser eficientes.

En este capítulo examinaremos el concepto económico de eficiencia técnica y de eficiencia asignativa. A continuación, se exponen las principales características del análisis envolvente de datos, herramienta matemática que utilizaremos para su medición. Por último se aborda paso a paso y de manera sencilla el uso del programa gratuito DEAP 2.1 para su utilización en casos prácticos.

9.2. EL CONCEPTO DE EFICIENCIA PRODUCTIVA

Cualquier proceso productivo se caracteriza por utilizar una cantidad de factores productivos a los que llamaremos *inputs* para obtener una producción que denominaremos *output* dada una tecnología disponible. Los dos inputs más utilizados son en su acepción más amplia el trabajo y el capital. El trabajo a su vez puede estar dividido en trabajadores de alta y de baja cualificación, por nivel de estudios o por alguna otra división que tenga en cuenta el grado de capital humano del trabajador. Por otro lado se entiende por capital el presupuesto (distinto al gasto en personal) de la unidad productiva dedicado a gastos corrientes, de maquinaria e infraestructuras necesario para llevar a cabo la actividad. El output además de recoger la dimensión cuantitativa de los resultados también debe recoger en aquellos casos que así lo requieran una medida cualitativa de estos resultados. Volviendo a la información a familias no sólo es necesario cuantificar las peticiones de información atendidas sino que también sería interesante evaluar el grado de satisfacción con la información recibida, la amabilidad en el trato, etc.

Volviendo a la definición del concepto de eficiencia, la mayor parte de los economistas entienden que este concepto tiene básicamente dos dimensiones:

La **eficiencia técnica**, como su propio nombre indica, es un concepto tecnológico que intenta analizar los procesos productivos y la organización de tareas, fijando su atención en las cantidades de factores productivos o inputs utilizadas y no en los costes o precios de los mismos. Puede expresarse tanto en términos de indicadores de resultados o outputs como en términos de factores productivos o inputs. En el primer caso, indicaría cuál es el logro del máximo producto(s) o servicio(s) posible(s) que se puede alcanzar para una combinación de factores productivos. En el segundo, indica la cantidad mínima de inputs requerida combinados en una determinada proporción para alcanzar un nivel dado de producto o de servicio.

Suponiendo que se ha logrado la eficiencia técnica, la **eficiencia asignativa** implica alcanzar el coste mínimo de producir un nivel dado de producto cuando se modifican las proporciones de los factores de producción utili-

zados de acuerdo con sus precios. Alternativamente, se puede definir como la obtención de una cantidad máxima de producto manteniendo el coste a través del reajuste de los factores de producción según sus costes de uso. Esto es, si existen múltiples combinaciones de factores productivos —trabajo y capital— para alcanzar un nivel de producción, la asignación asignativamente eficiente sería la elección de la tecnología más barata.

En definitiva, en la medición de la eficiencia técnica se parte de una proporción concreta de factores. Esta proporción puede variar si, por ejemplo, se utiliza una tecnología distinta pero no por precios o productividades como ocurre con la eficiencia asignativa.

Una forma alternativa de definir la eficiencia técnica y asignativa sería la siguiente. En el caso de la eficiencia técnica, figura 9.1, nos encontraríamos ante un proceso de producción caracterizado por una tecnología que requiere la utilización de dos inputs (trabajo - L y capital - K) para la obtención de un único output Y. Estaríamos actuando eficientemente desde el punto de vista técnico cuando nos encontráramos en un punto sobre la isocuanta [1] que caracteriza la tecnología frontera (puntos B y D). Alternativamente, se define una situación como eficiente asignativamente, cuando siéndolo desde el punto de vista técnico, estamos empleando la menor cantidad de recursos posible, dados los precios de los factores productivos (punto D). Nos encontramos sobre la curva isocoste [2] más baja (CC').

En la figura 9.1 se recoge la situación de una unidad productiva que para producir una cantidad de output utiliza la combinación de factores productivos representada por el punto A [3]. Suponiendo que la función de producción es conocida y que existen rendimientos constantes a escala la isocuanta no es más que las combinaciones de factores productivos necesarios para producir una cantidad dada de producción.

Así, dada la situación descrita en la figura anterior podemos deducir que la unidad productiva A no es técnicamente eficiente ya que la isocuanta indica que el output producido por la entidad A podría ser obtenida uti-

---

[1] Dada una función de producción Y = F(L,K), recibe el nombre de isocuanta las infinitas combinaciones lineales de trabajo y capital capaces de producir una cantidad dada de producción.

[2] La restricción presupuestaria de una unidad productiva viene caracterizada por la ecuación C = wL + rK donde w es el precio del factor productivo trabajo (salario) y r el precio del factor productivo capital (el tipo de interés). La isocoste representa las infinitas combinanciones de trabajo y capital posibles para gastar un presupuesto dado. Así, la isocoste más cercana al origen de coordenadas es la que requiere un presupuesto menor.

[3] Obsérvese que para representar gráficamente dos inputs (trabajo –L y capital –K) y un output (Y) hemos dividido los inputs por la cantidad de producción de cada unidad productiva. Alternativamente, la figura 8.1 podría ser representada con las cantidades absolutas de trabajo y capital asumiendo que todos los puntos pertenecientes y por encima de la isocuanta producen una unidad (isocuanta unitaria).

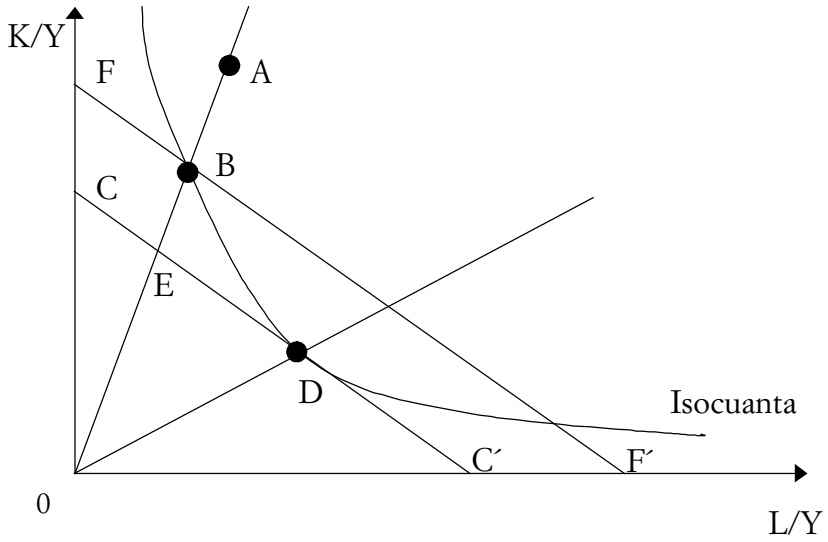


FIGURA 9.1. *Eficiencia Técnica y Eficiencia Asignativa*

lizando una proporción OB/OA de los inputs que realmente utiliza y sin variar la combinación de los mismos. Farrell (1957) define el cociente OB/OA como el índice de eficiencia técnica de la unidad productiva A. Observamos como la medida de eficiencia técnica de Farrell tomará el valor 1 si la entidad es técnicamente eficiente y valores más próximos a cero cuanto más ineficiente sea la entidad valorada.

La figura 9.1 permite también establecer la medida de la eficiencia precio. Si se supone que los precios de los factores productivos están representados por la pendiente de las rectas FF' y CC', el punto D es aquel en que permite obtener la misma producción que en B (está sobre la curva isocuanta) pero que minimiza el coste. Tanto B como D son eficientes técnicamente por estar situados sobre la isocuanta, pero los costes de producción en D son tan sólo una fracción OE/OB de los costes de producción en B y por eso Farrell considera al cociente OE/OB como la eficiencia asignativa de la unidad productiva situada en B. Esta medida de eficiencia asignativa, que Farrell refiere al punto B, mide también la eficiencia asignativa de la entidad A que estamos evaluando. En efecto, la eficiencia asignativa de A mide exclusivamente el exceso de costes en que se está incurriendo por combinar los inputs de una forma diferente a la óptima. Esto indica que la valoración de la eficiencia asignativa pura precisa la eliminación de la eficiencia técnica lo que, en nuestro caso equivale a situar la entidad A en el punto B y valorar el exceso de costes en esa situación, representada por el salto de B a D.

Farrell continua definiendo una medida de eficiencia global como el producto de la eficiencia técnica y la eficiencia asignativa.

$$\frac{OE}{OA} = \frac{OB}{OA} \times \frac{OE}{OB}$$

Siendo (OE/OA) la eficiencia Global, (OB/OA) la eficiencia técnica y (OE/OB) la eficiencia asignativa.

El análisis efectuado a partir de la figura 9.1 parte del conocimiento de la función de producción, representada por medio de la isocuanta, la cual constituye, como se ha visto, el punto de referencia para llevar a cabo las mediciones. Las situaciones reales, sin embargo, no se suelen caracterizar por el conocimiento de esas relaciones tecnológicas, lo cual complica la medición operativa de la eficiencia.

9.3. LA MEDICIÓN EMPÍRICA DE LA EFICIENCIA

La mayor parte de procesos productivos reales son desconocidos y por ello no se dispone automáticamente de la isocuanta para medir la eficiencia. Este problema fue también considerado por Farrell, quien propone un método original de estimación de la función de producción a partir de los datos de las entidades implicadas en la evaluación. Es justamente la estimación de lo que él denomina una función de producción empírica lo que ha tenido una mayor repercusión en la literatura posterior sobre la medición de la eficiencia en procesos productivos reales.

Para explicar la propuesta de Farrell nos serviremos de nuevo del análisis gráfico. En la figura 9.2 se representan, mediante puntos, las combinaciones de factores utilizadas por diferentes unidades productivas para obtener una cantidad de output.

Farrell impone dos condiciones a la isocuanta que va a estimar: que sea convexa y que no tenga en ningún punto pendiente positiva. La primera condición, normalmente planteada en teoría económica, significa que si dos puntos se pueden alcanzar en la práctica, entonces también se podrá obtener cualquier otro que sea una combinación lineal (media ponderada) de aquellos. La segunda condición, por su parte, es necesaria para asegurar que el aumento de los factores utilizados no implicará nunca una reducción en la cantidad de producto.

A partir de estas condiciones es fácil deducir que la isocuanta eficiente está representada en la figura 9.2 por la curva Y_0, es decir, por el conjunto de

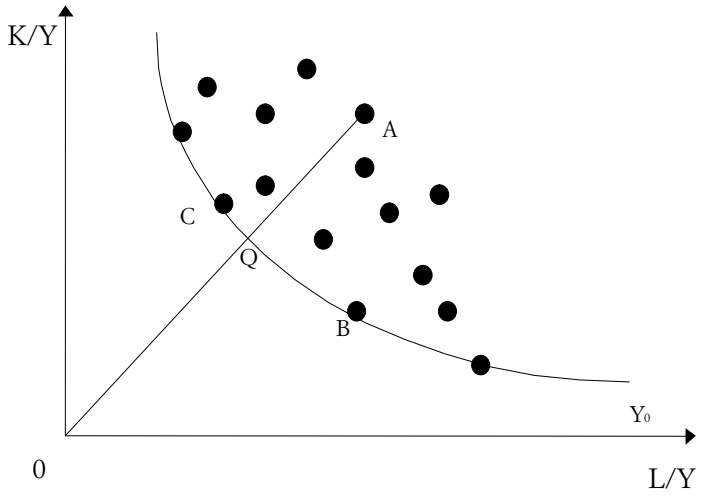


FIGURA 9.2. *La medición empírica del concepto de eficiencia*

puntos más próximos al origen, las unidades más eficientes en términos relativos, que puedan ser unidos a través de una curva convexa que no tenga en ningún punto pendiente positiva. Determinada la isocuanta eficiente Y_0, el proceso de medir la eficiencia de cualquier unidad productiva es el que hemos especificado en el comentario de la figura 9.1. Como allí destacábamos, se trata de comparar cada entidad que no pertenece a la isocuanta con otra entidad eficiente que utilice los factores productivos en la misma proporción (esto es, que se encuentre en el mismo radio vector desde el origen). En general, esa comparación se realizará con unidades hipotéticas que, empleando los factores en la misma proporción, se encuentran, sobre la isocuanta eficiente, pero que no se corresponden con ninguna observación real. Así, en la figura anterior, la eficiencia del punto A se mide comparando los factores que utiliza con los que usa la unidad ficticia Q que se construye como combinación lineal de las unidades reales eficientes B y C. Farrell señala que la esencia de su propuesta radica precisamente en la construcción de esas unidades hipotéticas y no en la representación de la isocuanta.

9.4. EL ANÁLISIS ENVOLVENTE DE DATOS (DEA) [4]

Formulación Básica

El modelo DEA fue desarrollado por Charnes, Cooper y Rhodes (1978). Mediante la utilización de técnicas de programación lineal, DEA

[4] Siglas de la expresión inglesa *Data Envelopment Analysis* por la que es comúnmente conocida y que utilizaremos en este capítulo.

compara la eficiencia técnica de un conjunto de unidades que producen outputs similares a partir de un conjunto común de inputs.

La eficiencia de la unidad que se pretende evaluar se define como la ratio de la suma ponderada de *outputs* con respecto a la suma ponderada de *inputs*. El problema matemático a resolver es:

$$\text{Max } h_0 = \frac{\sum_{r=1}^{s} U_r Y_{r0}}{\sum_{i=1}^{m} V_i X_{i0}}$$

Sujeto a:

$$\frac{\sum_{r=1}^{s} U_r Y_{rj}}{\sum_{i=1}^{m} V_i X_{ij}} \leq 1$$

$$U_r, V_j \geq 0 \; ; \; r = 1...s \; ; \; i = 1...m$$

Donde:

Y_{r0} = Cantidad de output r producido por la unidad evaluada.
X_{i0} = Cantidad de input i consumido por la unidad evaluada.
Y_{rj} = Cantidad de output r producido por la unidad j.
X_{ij} = Cantidad de input i consumido por la unidad j.
U_r = Ponderación asignada al output r.
V_i = Ponderación asignada al input i.

El problema fraccional así formulado consiste en *encontrar el conjunto de ponderaciones que maximizan el valor de los outputs de la unidad analizada con respecto a sus inputs, con la restricción de que aplicando estas mismas ponderaciones a las restantes unidades, ninguna debiera tener una relación output/input mayor que uno.* Si, sujeto a esta restricción, fuera factible encontrar un conjunto de ponderaciones con las que la ratio de eficiencia de la unidad productiva evaluada sea igual a 1, entonces será considerada eficiente. En caso contrario la unidad será considerada ineficiente, ya que incluso con el conjunto de ponderaciones más favorable puede encontrarse una ratio de eficiencia mayor.

Dado que el problema de optimización matemática anterior tiene infinitas soluciones en su versión fraccional (cualquier combinación lineal λU, λV también es solución del problema), el modelo puede transformarse fácilmente en un programa lineal para facilitar su resolución. Para ello, basta con maximizar el numerador de la función objetivo manteniendo constante el denominador:

$$\text{Max } h_0 = \sum_{r=1}^{s} u_r y_{r0}$$

$$\text{s.a.} \quad \sum_{i=1}^{m} v_i x_{i0} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0$$

$$v_i, u_r \geq 0$$

$j = 1,...,n$ r = 1,..., s i = 1,....,m

El programa lineal selecciona las ponderaciones que maximizan el output virtual de la unidad ($u_r\ y_{r\,0}$), condicionadas a que su input virtual ($v_i\ x_{i\,0}$) sea igual a la unidad, así como que la aplicación de dichas ponderaciones al resto de unidades de decisión no permita que su output virtual exceda del input virtual. La unidad será eficiente si su output virtual es unitario.

No obstante, en la práctica y por motivos computacionales, el cálculo de los índices de eficiencia resulta más sencillo si se utiliza la forma dual de este programa. La formulación dual en términos de maximización del output es la siguiente:

$$\text{Max}_{\theta,\lambda}\ \theta_0$$

$$\text{sujeto a} \quad \sum_{j=1}^{n} \lambda_j y_{rj} \geq \theta y_{r0}$$

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{i0}$$

$$\lambda_j \geq 0$$

$j = 1,...,n$ r = 1,..., s i = 1,....,m

En este caso, si q =1, la unidad evaluada se considerada eficiente, pues no existe otra que produzca más o que consiga el mismo nivel de producción con menores recursos que ella. El modelo también puede ser planteado como la minimización de inputs productivos dado un nivel de resultados.

$$\begin{aligned} &\text{Min}_{\theta,\lambda}\ \theta_0 \\ &\text{sujeto a} \quad \theta_0 x_{i0} - \sum_{j=1}^{N} \lambda_j x_{ij} \geq 0 \quad i \in \{1,\ldots,M\} \\ &\qquad -y_{r0} + \sum_{j=1}^{N} \lambda_j y_{rj} \geq 0 \quad r \in \{1,\ldots,S\} \\ &\qquad \theta_0, \lambda_j \geq 0 \end{aligned}$$

Donde *i* denota input, *r* denota output y *j* unidad productiva. Los modelos anteriores describen un proceso productivo con rendimientos constantes a escala, esto es, se asume que si todos los inputs aumentaran en un porcentaje el output también aumentaría en ese porcentaje. Esta hipótesis fue relajada por Banker, Charnes y Cooper (1984) extendiendo el modelo anterior al caso de rendimientos variables a escala. Bajo esta formulación entendemos que la tecnología no impone aumentos equiproporcionales de inputs y outputs. Para ello es necesario añadir una restricción a los modelos anteriores (tanto en la versión orientada al output como al input) con el fin de establecer la convexidad de la frontera productiva.

$$\begin{aligned} &\text{Min}_{\theta,\lambda}\ \theta_0 \\ &\text{sujeto a} \quad \theta_0 x_{i0} - \sum_{j=1}^{N} \lambda_j x_{ij} \geq 0 \quad i \in \{1,\ldots,M\} \\ &\qquad -y_{r0} + \sum_{j=1}^{N} \lambda_j y_{rj} \geq 0 \quad r \in \{1,\ldots,S\} \\ &\qquad \sum_{j=1}^{N} \lambda_j = 1 \\ &\qquad \theta_0, \lambda_j \geq 0 \end{aligned}$$

Normalmente las aplicaciones relacionadas con evaluación del sector público o instituciones sin ánimo de lucro suelen emplear una formulación de maximización del output bajo rendimientos variables a escala aunque en cada caso concreto se deberá elegir la especificación más adecuada.

Eficiencia de escala

Otro concepto importante es el de la eficiencia de escala. La eficiencia de escala relaciona el tamaño de la unidad productiva con su productividad. Se calcula como el cociente entre la eficiencia con rendimientos

constantes y variables a escala. Así, la figura 9.3 muestra como una unidad puede ser eficiente (G y J) y sin embargo no operar en la escala óptima (H). La unidad G es técnicamente eficiente pues pertenece a la frontera productiva. Sin embargo, opera en un contexto de rendimientos crecientes a escala. Esto es si G aumentara su input su output aumentaría en mayor proporción. De forma análoga la unidad J opera bajo rendimientos decrecientes a escala. Si disminuyera su input su producción disminuiría en un porcentaje menor.

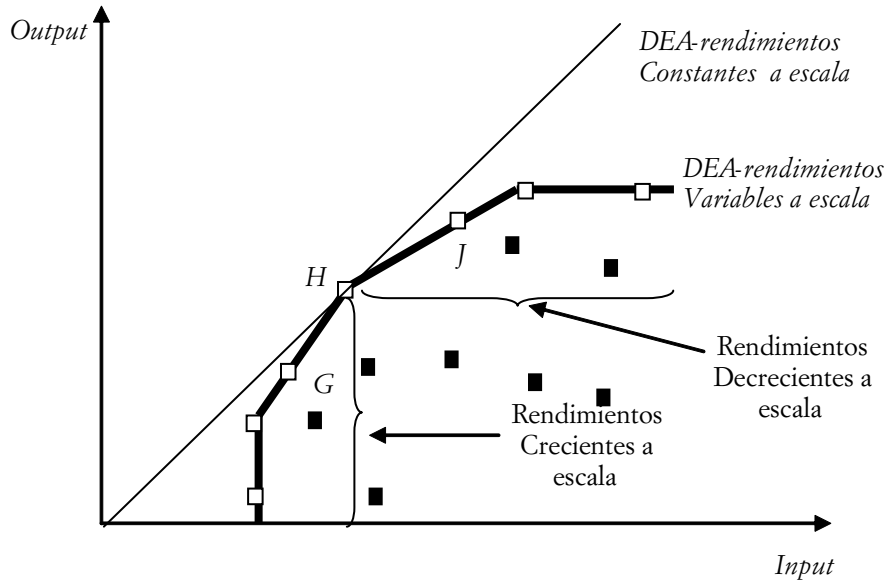


FIGURA 9.3. *Eficiencia de escala*

Ineficiencia no radial o variables de holgura

Además de ofrecer un índice que refleja el porcentaje de incremento de outputs (o reducción de inputs) necesario para que una unidad sea eficiente, el DEA también permite detectar posibles reducciones adicionales en los inputs o incrementos potenciales en los outputs mediante la incorporación al modelo dual de las denominadas variables de holgura [5].

En la orientación input estas holguras representan la cantidad que se podría ahorrar cada productor en la utilización de los mismos aún en el caso de ser eficiente, mientras que, en la orientación output se interpretan con cuánto se podría incrementar la producción aunque se pertenezca a la frontera.

[5] A menudo los trabajos de eficiencia se refieren a esta medida utilizando la voz inglesa *slack*.

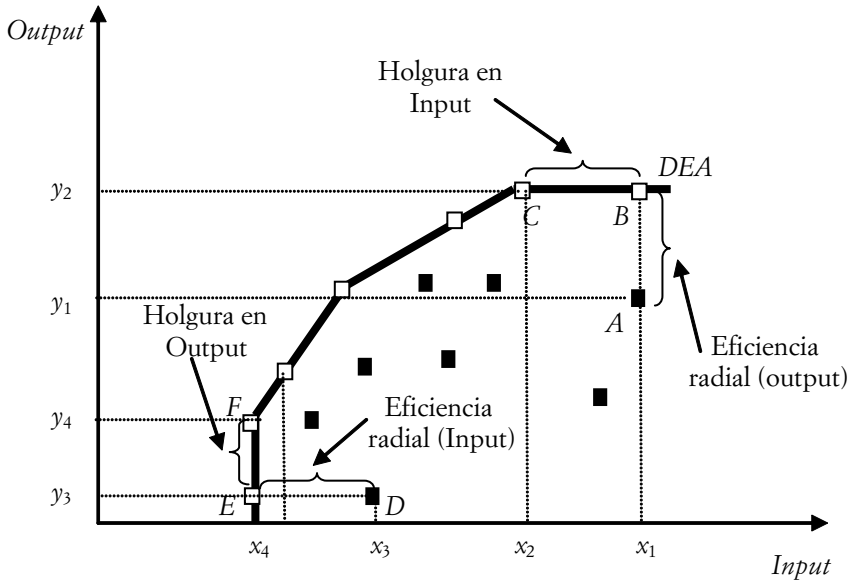


FIGURA 9.4. *Representación gráfica del concepto de holgura*

La Figura 9.4 nos permite ilustrar esta definición. En ella se muestran los resultados de eficiencia radial y no radial obtenidos por dos unidades A y D que utilizan un input para producir un output.

Así, la unidad A es ineficiente pues no pertenece a la frontera productiva. Para pertenecer a la frontera deber aumentar su producción desde y_1 a y_2 permaneciendo su cantidad de input x_1 constante hasta situarse en el punto B. ¿Es B una unidad eficiente? En principio podemos pensar que B es eficiente dado que pertenece a la frontera productiva. Sin embargo el gráfico también muestra como una vez en B la tecnología permite reducir el input consumido desde x_1 a x_2 sin que la producción y_2 varíe. Decimos en este caso que el punto B presenta una holgura en el input. De manera análoga se puede analizar la situación de la unidad D. La unidad D es ineficiente técnicamente y para ser eficiente su proyección radial en términos de input le llevaría a reducir su éste desde x_3 a x_4 hasta situarse en el punto E. Al igual que en el caso anterior el punto E presenta una holgura en el output ya que con una cantidad de x_4 de input la tecnología permite producir y_4 por lo que el punto E no es eficiente.

Estas variables de holgura se pueden incluir en el DEA a través de las siguientes expresiones:

$$s_i^- = x_{i0} - \left(\sum_{j=1}^{n} \lambda_j x_{ij}\right) \qquad s_r^+ = \left(\sum_{j=1}^{n} \lambda_j y_{rj}\right) - \theta y_{r0}$$

donde s_i^- refleja el exceso adicional del input i s_r^+ y el aumento potencial del output r. De este modo, el modelo dual de maximización del output que vimos con anterioridad bajo rendimientos constantes a escala puede ser extendido adoptando la siguiente forma:

$$\text{Max} \quad \theta+\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

$$\text{sujeto a} \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{i0} \qquad i=1,2,\ldots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = \theta y_{r0} \qquad r=1,2,\ldots,s$$

$$\lambda_j \geq 0;\ s_r^+ \geq 0;\ s_i^- \geq 0 \quad j=1,2,\ldots,n$$

en el que q_0 es el índice de eficiencia, son las ponderaciones y s_i^- y s_r^+ son las variables de holgura de los inputs y outputs, respectivamente. En este caso, un productor es relativamente eficiente si y sólo si su índice de eficiencia es igual a la unidad y además todos las holguras son nulas.

Modelos DEA de producción pura

En el ámbito de la investigación social de la familia es frecuente que todas las unidades de decisión tengan el mismo presupuesto. Ejemplo de ello serían equipos de trabajo o profesionales que en mismo número y con los mismos recursos atienden a diversos colectivos. Lovell y Pastor (1999) demuestran que la formulación de este tipo de problema, en el que todas las unidades utilizan la misma cantidad de input, en un programa DEA orientado al output en sus versiones de rendimientos constantes y variables a escala coinciden [6]. A partir de este resultado el programa que resolveremos toma la siguiente forma.

[6] Como señalan Lovell y Pastor (1999) una consecuencia inmediata de este resultado es que en este tipo de modelos «de producción pura» los rendimientos a escala de todas las unidades de decisión son siempre constantes.

$$\text{Max}\left[\theta_0 + \varepsilon\left(\sum_{r=1}^{S} s_{r0}^{-}\right)\right]$$

Sujeto a

$$\sum_{j=1}^{N} \lambda_j y_{rj} - s_{r0}^{-} = \theta_0 y_{r0} \quad r \in \{1,\ldots,S\}$$

$$\sum_{j=1}^{N} \lambda_j = 1$$

$$\theta_0, \lambda_j, s_{r0}^{-} \geq 0$$

Donde θ_0 es el nivel de eficiencia de la unidad analizada, y_{rj} es el output r de la unidad productiva j y λ es el vector de intensidad (1 x N) que pondera la actividad de los distintos procesos productivos observados. En esta formulación ε es un número pequeño y positivo y s_{r0}^{-} representa las holguras en los outputs. En este modelo sólo existe un único input cuyo valor es igual a 1 para todas las unidades o, lo que es lo mismo, se considera que todas las unidades se enfrentan a la tarea de maximización de los resultados con los mismos inputs.

9.5. EL PROGRAMA DEAP (DATA ENVELOPMENT ANALYSIS PROGRAM)

No sólo es posible utilizar Excel para la resolución de problemas de programación lineal matemática como DEA sino que además éste es frecuentemente utilizado en investigación [7]. Esta utilidad está incluida en Herramientas → Solver que permite la programación directa de estos modelos. Como excepción al resto del libro en este capítulo no utilizaremos Excel como herramienta de trabajo. Las razón fundamental es que la programación en Solver va más allá de los objetivos meramente aplicados de esta obra. En su lugar se propone utilizar el software gratuito DEAP 2.1 desarrollado por el profesor Tim Coelli y que puede ser descargado en:

http://www.uq.edu.au/economics/cepa/deap.htm

El archivo DEAP.pdf contiene las instrucciones detalladas de uso del programa por lo que a continuación tan solo se exponen algunas líneas que resumen su funcionamiento [8].

---

[7] El lector interesado en estas herramientas puede acudir por ejemplo a http://www.deafrontier.com o bien a http://www.prodtools.com/Products.html#xldea. Estas herramientas pueden ser adquiridas a un coste razonable. Otro software gratuito que puede utilizarse para calcular eficiencia es el Efficiency Measurement System (EMS) http://www.wiso.uni-dortmund.de/lsfg/or/scheel/ems/

[8] Se remite al lector a que consulte directamente estas instrucciones.

Para la ejecución del programa DEAP 2.1 necesitamos básicamente 5 ficheros:

- EL ejecutable DEAP.EXE
- El fichero con los parámetros básicos para inicializar el algoritmo DEAP.000
- Un fichero con los datos del problema *.dta aunque puede usarse cualquier otro archivo de texto por ejemplo con extensión *.txt
- Un fichero con las instrucciones que damos al programa *.ins
- Finalmente el programa genera un fichero de salida *.out

La carpeta con el programa proporciona ficheros de ejemplo que pueden ser copiados y usados para nuestros propósitos reemplazando la información que contienen por la de nuestro problema.

Ya hemos visto que el programa requiere que los datos estén contenidos en un fichero de texto, desde EXCEL los datos pueden ser guardados como un fichero de texto y deben ser dispuestos de tal forma que la información aparezca siguiendo cierto orden.

- Los datos deben aparecer por observaciones, una fila para cada unidad productiva.
- En las columnas los valores de los outputs deben ir en primer lugar, y a continuación irán los inputs (de izquierda a derecha).
- Los datos no deben contener ni el nombre de las variables ni de las unidades.
- Los datos deben utilizar puntos como separadores de decimales en lugar de comas.

El fichero EG1.DTA contiene 5 observaciones, 1 output y dos inputs. El output, tal y como se ha comentado con anterioridad, está en la primera columna y los inputs en las dos siguientes (tabla 1).

TABLA 1

1	2	5
2	2	4
3	6	6
1	3	2
2	6	2

El fichero EG1.INS contiene la descripción de los datos y las instrucciones que daremos (tabla 2).

TABLA 2

C:\eg1.dta	DATA FILE NAME
C:\eg1.out	OUTPUT FILE NAME
5	NUMBER OF FIRMS
1	NUMBER OF TIME PERIODS
1	NUMBER OF OUTPUTS
2	NUMBER OF INPUTS
0	0=INPUT AND 1=OUTPUT ORIENTATED
0	0=CRS AND 1=VRS
0	0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA, 3=DEA(1-STAGE), 4=DEA(2-STAGE)

Las primeras dos líneas contienen la ruta en nuestro ordenador donde se encuentran los datos y donde queremos que nos guarde los resultados (extensión *.OUT). Si no establecemos una ruta el programa entenderá que tanto los datos y los resultados estarán en la misma carpeta del programa. A continuación nombramos el número de unidades productivas (5), el número de períodos (1), el número de outputs (1) y de inputs (2), la orientación del problema (input), la escala en la que se opera (0 = rendimientos constantes a escala) y el algoritmo DEA que usaremos (0 es el estándar).

Para correr el programa debemos hacer doble-click en el ejecutable DEAP.EXE. A continuación se nos pide en una pantalla bajo entorno DOS que introduzcamos la ruta en la que se encuentra el fichero de instrucciones. El fichero con los resultados es enviado a la ruta que hemos programado.

EJERCICIO 9.1. En una gran ciudad existen 21 Centros de Ayuda a la Familia (CAF) que atienden fundamentalmente tres funciones: información a las familias, actividades de formación y servicio de mediación. Para ello cada una de ellas cuenta con un presupuesto sobre el que tiene un grado de discrecionalidad elevado. La información está contenida en el archivo CAF-TEMA9.xls

Se pide:

a) **Medir la eficiencia técnica de las CAF bajo rendimientos variables a escala y utilizando una orientación al output.**

b) **Los objetivos en la frontera para la CAF 18**

c) **Detectar los centros eficientes que más veces se utilizan para calcular los objetivos de las CAF ineficientes a fin de aprender de sus buenas prácticas.**

d) **Repetir los resultados bajo rendimientos constantes a escala. Comente como varían los resultados**

El fichero CAF-TEMA9.xls presenta la información original que se muestra en la siguiente tabla:

Datos del problema

	OUTPUT1	*OUTPUT2*	*OUTPUT3*	*INPUT1*
	Información	*Formación*	*Mediación*	*Presupuesto*
CAF1	21237	4840	8486	627839
CAF2	12732	4275	9144	507557
CAF3	5840	5100	4825	414631
CAF4	6249	4000	4245	491052
CAF5	8088	4900	3200	253331
CAF6	8325	4200	4320	707871
CAF7	3455	1500	4000	183797
CAF8	16988	6600	1100	524695
CAF9	19156	2800	6200	558864
CAF10	35276	9000	17000	1291658
CAF11	29399	5400	8357	524443
CAF12	25435	3500	3795	737165
CAF13	25060	23400	25264	508901
CAF14	22013	5000	1000	635147
CAF15	19545	5100	2500	370720
CAF16	15133	7410	3700	275270
CAF17	19476	4800	4615	1315347
CAF18	9160	2800	3500	258240
CAF19	5509	2350	6050	522432
CAF20	20866	8000	18000	968703
CAF21	9123	1730	1916	497667

Para utilizar los datos con el programa DEAP debemos eliminar las etiquetas de fila y columna y guardar la información como texto delimitado por tabulaciones (txt) en el archivo caftexto. A continuación debemos establecer los comandos del fichero de instrucciones. Para ello utilizaremos el fichero caf.txt

Instrucciones en el fichero caf.txt

caftexto.txt	DATA FILE NAME
caf.out	OUTPUT FILE NAME
21	NUMBER OF FIRMS
1	NUMBER OF TIME PERIODS
3	NUMBER OF OUTPUTS
1	NUMBER OF INPUTS
1	0=INPUT AND 1=OUTPUT ORIENTATED
1	0=CRS AND 1=VRS
0	0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA, 3=DEA(1-STAGE), 4=DEA(2-STAGE)

Una vez establecidas las instrucciones se guarda el fichero de instrucciones y de datos en la misma carpeta del programa ya que no hemos establecido ruta. A continuación se ejecuta el programa en una pantalla como la de la figura 9.5.

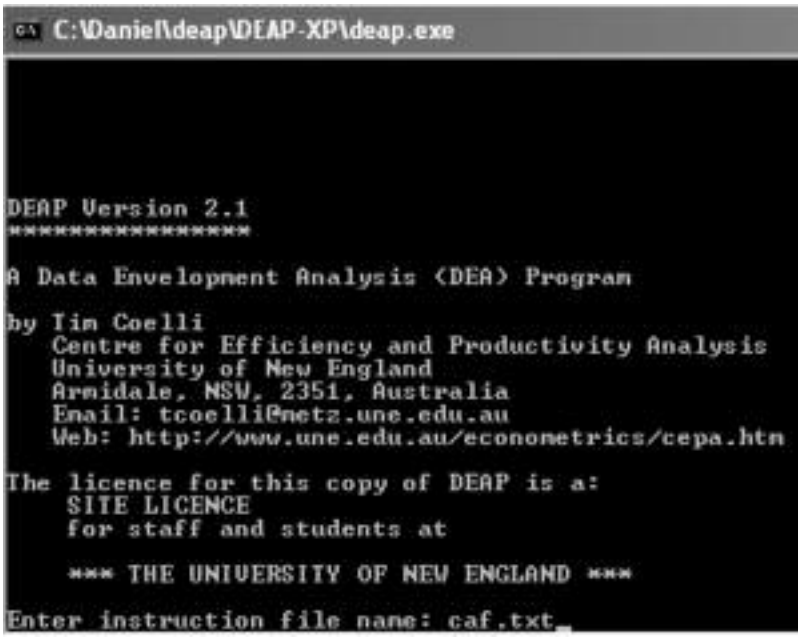


FIGURA 9.5. *Ejecución del programa DEAP*

Una vez ejecutado el programa genera los resultados en el fichero caf.out que puede ser abierto con el Bloc de Notas o con cualquier otro editor de texto. Los principales resultados son los siguientes.

EFFICIENCY SUMMARY:

firm	crste	vrste	scale	
1	0.617	0.717	0.860	drs
2	0.485	0.485	1.000	-
3	0.282	0.300	0.941	irs
4	0.245	0.246	0.997	drs
5	0.598	0.804	0.744	irs
6	0.220	0.288	0.765	drs
7	0.438	1.000	0.438	irs
8	0.585	0.604	0.968	drs
9	0.616	0.650	0.948	drs
10	0.506	1.000	0.506	drs
11	1.000	1.000	1.000	-
12	0.616	0.820	0.751	drs
13	1.000	1.000	1.000	-
14	0.620	0.733	0.846	drs
15	0.945	0.949	0.996	irs
16	1.000	1.000	1.000	-
17	0.265	0.552	0.480	drs
18	0.646	0.753	0.857	irs
19	0.233	0.239	0.974	drs
20	0.427	0.766	0.557	drs
21	0.327	0.327	0.999	-
mean	0.556	0.678	0.839	

La tabla anterior muestra los resultados con rendimientos constantes y variables a escala. En nuestro caso estamos interesados en la eficiencia con rendimientos variables por lo que podemos comprobar como las CAF 7, 10, 11, 13 y 16 son eficientes siendo el resto ineficientes. Asimismo se muestra la eficiencia de escala que señala que independientemente de sus eficiencias técnicas las CAF 2, 11, 13, 16 y 21 operan en la escala adecuada. A continuación se muestran los resultados de las holguras en inputs y outputs.

SUMMARY OF OUTPUT SLACKS:

firm output:	1	2	3
1	0.000	1582.683	0.000
2	0.000	7964.078	0.000
3	0.000	0.000	2219.999
4	0.000	0.000	664.099
5	0.000	0.000	2341.256
6	0.000	0.000	3291.455
7	0.000	0.000	0.000
8	0.000	0.000	11757.447
9	0.000	2092.795	0.000
10	0.000	0.000	0.000
11	0.000	0.000	0.000
12	0.000	2128.459	6123.833
13	0.000	0.000	0.000
14	0.000	0.000	9093.068
15	0.000	1265.312	2849.276
16	0.000	0.000	0.000
17	0.000	305.977	8641.059
18	0.000	2627.220	0.000
19	2055.144	13586.711	0.000
20	0.000	9881.217	0.000
21	0.000	331.750	2004.186
mean	97.864	1988.867	2332.651

SUMMARY OF INPUT SLACKS:

firm input:	1
1	0.000
2	0.000
3	0.000
4	0.000
5	0.000
6	0.000
7	0.000
8	0.000
9	0.000
10	0.000
11	0.000
12	0.000
13	0.000
14	0.000
15	0.000
16	0.000
17	23689.000
18	0.000
19	13531.000
20	292676.094
21	0.000
mean	15709.338

SUMMARY OF OUTPUT TARGETS:

firm output:	1	2	3
1	29619.454	8333.079	11835.508
2	26262.167	16782.077	18861.236
3	19487.877	17018.522	18320.856
4	25439.077	16283.615	17945.085
5	10060.080	6094.757	6321.505
6	28865.584	14562.817	18270.352
7	3455.000	1500.000	4000.000
8	28111.905	10921.743	13577.737
9	29466.906	6399.923	9537.211
10	35276.000	9000.000	17000.000
11	29399.000	5400.000	8357.000
12	31028.487	6398.155	10753.403
13	25060.000	23400.000	25264.000
14	30036.185	6822.374	10457.543
15	20597.836	6640.035	5483.944
16	15133.000	7410.000	3700.000
17	35276.000	9000.000	17000.000
18	12163.975	6345.466	4647.807
19	25060.000	23400.000	25264.000
20	27241.211	20325.466	23499.559
21	27865.983	5615.994	7856.561

SUMMARY OF INPUT TARGETS:

firm input:	1
1	627839.000
2	507557.000
3	414631.000
4	491052.000
5	253331.000
6	707871.000
7	183797.000
8	524695.000
9	558864.000
10	1291658.000
11	524443.000
12	737165.000
13	508901.000
14	635147.000
15	370720.000
16	275270.000
17	1291658.000
18	258240.000
19	508901.000
20	676026.906
21	497667.000

Resultados detallados para la CAF 18

Results for firm: 18

Technical efficiency = 0.753

Scale efficiency = 0.857 (irs)

PROJECTION SUMMARY:

variable		original value	radial movement	slack movement	projected value
output	1	9160.000	3003.975	0.000	12163.975
output	2	2800.000	918.246	2627.220	6345.466
output	3	3500.000	1147.807	0.000	4647.807
input	1	258240.000	0.000	0.000	258240.000

LISTING OF PEERS:

peer	lambda weight
16	0.672
7	0.288
13	0.040

La CAF 18 opera bajo rendimientos crecientes a escala con una ineficiencia técnica del 75,3%. Ello supone que para ser eficiente debe aumentar sus objetivos en un 24,7% (proyección radial). Además de la proyección radial en la frontera la unidad presenta una holgura en el output 2 lo que indica que es posible que realice aumentos adicionales a los de la proyección radial en este output sin variar su cantidad de input. Las CAF de referencia en la frontera para la CAF 18 son en este orden las CAF 16, 7, 13 cuya combinación lineal determinan el punto en la frontera donde se proyecta la CAF 18.

PEER COUNT SUMMARY:
(i.e., no. times each firm is a peer for another)

firm	peer count:
1	0
2	0
3	0
4	0
5	0
6	0
7	3
8	0
9	0
10	8
11	10
12	0
13	12
14	0
15	0
16	7
17	0
18	0
19	0
20	0
21	0

La tabla anterior muestra el recuento de veces en que las unidades eficientes sirven de referencia para las ineficientes. Las unidades 11 y 13 seguidas de la 10 y la 16 son las que más veces se utilizan para comparar respecto al resto y por tanto sería muy interesante estudiar como gestionan los recursos a fin de aprender y trasladar este conocimiento al resto de CAF.

EJERCICIO 9-2. En un pequeño país existen 26 equipos para el trabajo con menores y familias en distintas problemáticas. Sus funciones principales son conocer y supervisar la problemática y las necesidades de las mismas mediante reuniones con las autoridades, detectar casos nuevos y atender a todas las familias que lo requieran. Se sabe que todos los equipos son homogéneos en personal y presupuesto.

Se pide:

a) **Medir la eficiencia técnica de los equipos con un modelo de producción pura.**

El fichero Equipos-TEMA9.xls presenta la información original. Al tratarse de unidades que utilizan el mismo input debemos añadir una columna adicional con valor uno para todas las unidades.

Datos del problema

	Número de reuniones	*Casos vistos*	*Casos Nuevos*	*Input*
Equipo 1	45	460	121	1
Equipo 2	23	562	91	1
Equipo 3	37	441	143	1
Equipo 4	43	396	80	1
Equipo 5	25	480	108	1
Equipo 6	39	325	119	1
Equipo 7	36	443	93	1
Equipo 8	39	259	49	1
Equipo 9	37	394	31	1
Equipo 10	36	265	100	1
Equipo 11	24	422	73	1
Equipo 12	23	420	50	1
Equipo 13	32	337	102	1
Equipo 14	34	235	92	1
Equipo 15	34	256	69	1
Equipo 16	31	234	25	1
Equipo 17	28	259	71	1
Equipo 18	22	285	79	1
Equipo 19	26	176	74	1
Equipo 20	25	274	54	1
Equipo 21	24	199	50	1
Equipo 22	21	160	39	1
Equipo 23	20	110	33	1
Equipo 24	20	152	50	1
Equipo 25	20	91	43	1
Equipo 26	7	42	14	1

A continuación procedemos como en el ejercicio anterior. Eliminamos las etiquetas de las filas y las columnas y guardamos los datos con extensión txt. El fichero de instrucciones tomaría la siguiente forma

```
equipos.txt       DATA FILE NAME
equipos.out       OUTPUT FILE NAME
26        NUMBER OF FIRMS
1         NUMBER OF TIME PERIODS
3         NUMBER OF OUTPUTS
1         NUMBER OF INPUTS
1         0=INPUT AND 1=OUTPUT ORIENTATED
0         0=CRS AND 1=VRS
0         0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA,
3=DEA(1-STAGE), 4=DEA(2-STAGE)
```

Los resultados son análogos a los del ejercicio 9.1.

firm	te
1	1.000
2	1.000
3	1.000
4	0.956
5	0.945
6	0.924
7	0.912
8	0.867
9	0.846
10	0.813
11	0.798
12	0.788
13	0.776
14	0.758
15	0.756
16	0.689
17	0.622
18	0.611
19	0.594
20	0.583
21	0.533
22	0.467
23	0.444
24	0.444
25	0.444
26	0.156

SUMMARY OF OUTPUT SLACKS:

firm output:	1	2	3
1	0.000	0.000	0.000
2	0.000	0.000	0.000
3	0.000	0.000	0.000
4	0.000	45.581	37.279
5	2.814	0.000	0.000
6	0.000	101.749	0.000
7	0.000	0.000	11.499
8	0.000	161.154	64.462
9	0.000	0.000	82.637
10	0.000	132.358	0.000
11	0.000	0.000	9.141
12	0.000	0.000	35.975
13	0.000	16.916	0.000
14	0.000	149.606	0.000
15	0.000	121.176	29.676
16	0.000	120.323	84.710
17	0.000	43.750	6.893
18	0.000	0.000	0.000
19	0.000	160.942	0.000
20	0.000	0.000	25.495
21	0.000	86.875	27.250
22	0.000	117.143	37.429
23	0.000	212.500	46.750
24	0.000	118.000	8.500
25	0.000	255.250	24.250
26	0.000	190.000	31.000
mean	0.108	78.205	21.652

PEER COUNT SUMMARY:

(i.e., no. times each firm is a peer for another)

firm	peer count:
1	22
2	7
3	7
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0

SUMMARY OF OUTPUT TARGETS:

firm output:	1	2	3
1	45.000	460.000	121.000
2	23.000	562.000	91.000
3	37.000	441.000	143.000
4	45.000	460.000	121.000
5	29.264	507.857	114.268
6	42.189	453.324	128.730
7	39.466	485.656	113.454
8	45.000	460.000	121.000
9	43.745	465.821	119.288
10	44.276	458.281	122.990
11	30.092	529.119	100.671
12	29.202	533.247	99.457
13	41.221	451.025	131.392
14	44.860	459.667	121.385
15	45.000	460.000	121.000
16	45.000	460.000	121.000
17	45.000	460.000	121.000
18	36.016	466.573	129.331
19	43.735	456.997	124.478
20	42.871	469.869	118.097
21	45.000	460.000	121.000
22	45.000	460.000	121.000
23	45.000	460.000	121.000
24	45.000	460.000	121.000
25	45.000	460.000	121.000
26	45.000	460.000	121.000

El equipo 21 es referencia en 22 ocasiones y por tanto la unidad que más veces sirve al resto como ejemplo de unidad eficiente.

BIBLIOGRAFÍA

Banker, R. D.; Charnes, A. y Cooper, W. W. (1984). Models for estimating technical and scale efficiencies in data envelopment analysis, *Management Science* 30 (9), pp. 1078-92.

Coelli, T., Prasada Rao, D. S., O'Donnell, C. J. y Battese, G. E. (1998) An introduction to efficiency and productivity analysis. Second Edition. Springer. New York.

Charnes, A., Cooper, W. y Rhodes, E. (1978): «Measuring the Efficiency of Decision Making Units». *European Journal of Operational Research*. Vol. 2 (6), pp. 429-444.

Farrell, M.J. (1957): «The measurement of efficiency productive». *Journal of the Royal Statistical Society*, serie A,120, pp. 253-266.

Leibenstein, H. (1966): «Allocative efficiency and x-efficiency». *American Economic Review*, 56, pp. 392-495.

Lovell, C. A. K. y Pastor, J. T. (1999). «Radial DEA models without inputs or without outputs». *European Journal of Operational Research*, 118, 46-51.

CAPÍTULO X
DISEÑO EXPERIMENTAL

CÉSAR PÉREZ LÓPEZ
DANIEL SANTÍN GONZÁLEZ

10.1. CONTRASTE DE HIPÓTESIS

En la gestión de políticas de ayuda a la familia es fundamental conocer cuál ha sido el grado de éxito de las medidas llevadas a cabo. No es ocioso volver a recordar que los recursos son escasos y que el gasto o inversión de los mismos en políticas activas deben conseguir el máximo objetivo posible. Alguna de las preguntas que podemos hacernos tras una intervención social o económica; por ejemplo para reducir el alcoholismo en jóvenes, reducir la pobreza o incentivar la natalidad son: ¿Hubieran sido los resultados obtenidos distintos si no hubiésemos aplicado ninguna medida?, ¿Cuál de las políticas A, B o C aplicadas cada una en tres ayuntamientos similares está consiguiendo mejores resultados? Un análisis superficial de estas preguntas contestaría simplemente que la política con un resultado más alto es la mejor. Sin embargo para un análisis más detallado deberemos indagar acerca de si las diferencias obtenidas son o no estadísticamente significativas. En este marco surge el diseño experimental que tiene como objetivo dilucidar si existe evidencia para rechazar o no rechazar las preguntas anteriores formuladas en términos de hipótesis.

Una hipótesis estadística es simplemente una afirmación que se hace sobre una o más características de una población. Para contrastar la hipótesis que se formula acerca de la población, podríamos tomar todos y cada uno de los elementos de la población y ver si la afirmación que se hace es cierta o falsa. Sin embargo, esto quizá no sea posible realizarlo, bien porque la población en estudio sea infinita, bien porque la forma de llevar a cabo la investigación sería muy costosa. Entonces es necesario conformarse, a la hora de realizar la investigación que nos sirva para contrastar la afirmación, con observar una muestra aleatoria simple de la citada población (véase el capítulo 2).

Para la realización del contraste se utiliza un estadístico cuya distribución en el muestreo se conoce si la hipótesis que hemos hecho es verdadera. Extraída la muestra, el estadístico tomará un cierto valor, que o bien puede llevarnos a sospechar que la hipótesis no es razonable y debe ser rechazada, o, por el contrario, puede considerarse como justificación de la hipótesis. Sin embargo, tendremos siempre que tener presente que

tanto en un caso como en otro podemos equivocarnos, es decir, podemos rechazar una hipótesis siendo verdadera o bien aceptarla siendo falsa.

A lo largo de este capítulo examinaremos los principales contrastes estadísticos para evaluar la diferencia de medias en una variable cuantitativa que se produce cuando dos o más muestras son idénticas en todas las variables salvo en un factor.

Para trabajar con contrastes de hipótesis, es necesario tener muy claros como mínimo los siguientes conceptos:

- *Contraste de hipótesis*: procedimiento estadístico mediante el cual se investiga la aceptación o rechazo de una afirmación acerca de una o varias características de una población estadística.
- *Hipótesis nula, H_0*: es la hipótesis que se quiere contrastar, y es, por tanto, la que se acepta o rechaza como conclusión del contraste.
- *Hipótesis alternativa, H_1*: es la hipótesis que nos sitúa frente a H_0, de forma que si se acepta H_1 se rechaza H_0 y, recíprocamente, si se rechaza H_1, se acepta H_0.
- *Estadístico de contraste o función de decisión del contraste*: es una función de la muestra aleatoria simple cuya distribución se utiliza para el contraste.
- *Región crítica*: conjunto de valores del estadístico del contraste que lleva a la decisión de rechazar la hipótesis nula H_0.
- *Región de aceptación*: conjunto de valores del estadístico que lleva a la decisión de aceptar la hipótesis nula H_0.
- *Error del tipo I*: error que se comete en la decisión del contraste cuando se rechaza la hipótesis nula H_0, siendo correcta.
- *Error del tipo II*: error que se comete en la decisión del contraste cuando se acepta la hipótesis nula H_0, siendo falsa. Su probabilidad se denota por β.
- *Nivel de significación*: probabilidad de cometer el error del tipo *I*, lo denotaremos por α (también se le suele llamar tamaño de un contraste o grado de significación).
- *Potencia de un contraste*: probabilidad de rechazar la hipótesis nula H_0 siendo falsa. Se denota por $1-\beta$. (Utilizaremos siem-

pre contrastes de máxima potencia, dentro de los que tienen un determinado tamaño o nivel de significación).

- *Contraste unilateral*: aquél cuya región crítica está formada por un solo conjunto de puntos de la recta real.

- *Contraste bilateral*: aquél cuya región crítica está formada por dos conjuntos de puntos de la recta real disjuntos.

- *Curva de potencia*: curva que representa la potencia de un test $1-b$ en función de los verdaderos valores del parámetro (valores que toma el parámetro bajo la hipótesis nula).

- *Curva caracter stica de operac*: curva que representa el complementario de la potencia de un test b en función de los verdaderos valores del parámetro (valores que toma el parámetro bajo la hipótesis nula).

- *Factor*: Variable no métrica (característica de la población, política de familia seguida, etc.) que sirve para asignar a las observaciones a una de las categorías predefinidas en el factor que originan las diferencias en la variable estudiada.

En un contraste de hipótesis, nuestro comportamiento irá orientado, una vez fijado el nivel de significación de todas las regiones críticas con el mismo nivel de significación, a elegir aquella en la que la potencia del contraste sea la mayor (es decir, donde la probabilidad de cometer el error tipo II sea la menor). A esa región crítica de potencia máxima la denominaremos *región crítica de máxima potencia*, insistiendo en que su significado es el de ser aquella región del espacio muestral donde, para un nivel de significación dado, la probabilidad de incurrir en un error de segunda especie es mínimo. Según lo dicho, el procedimiento de contrastación de una hipótesis respecto de su alternativa requiere, una vez elegido el nivel de significación, determinar la región crítica de mayor potencia, o región crítica prepotente. Mediante el teorema de Neyman-Pearson, que se expone en cualquier manual de estadística intermedia, se obtienen este tipo de regiones críticas más potentes.

Procedimiento clásico para realizar un contraste de hipótesis

Es clásico considerar en la realización de contrastes de hipótesis las siguientes fases o etapas:

1. *Enunciado y determinación de las hipótesis nula y alternativa H_0 y H_1:* En este primer paso se determina el parámetro relevante en el proble-

ma analizado, así como cuáles son los valores de este parámetro que constituyen la hipótesis nula y la hipótesis alternativa. Se dice, en este caso, que se trata de un contraste paramétrico, ya que se refiere a un parámetro de la variable aleatoria en observación *X*. En los denominados contrastes no paramétricos, las hipótesis hacen referencia a la muestra seleccionada de la variable aleatoria en observación *X*, o bien a características de la población *X*, como puede ser su distribución de probabilidad en los llamados contrastes de bondad del ajuste. No obstante, en este tipo de contrastes no se supone un modelo concreto de distribución de probabilidad para la variable aleatoria *X*.

2. *Elección del nivel de significación deseado para el contraste:* Se trata ahora de fijar el valor a, o probabilidad de cometer el error de tipo I en el contraste. Este valor se suele fijar, en las aplicaciones, en 0,1, 0,05 o 0,01.

3. *Especificación del tamaño muestral*: Hay que expresar el tamaño *n* de la muestra en la cual se van a basar los resultados del contraste para toda la población. La muestra particular será del tipo $(x_1, x_2, \ldots, x_n)$.

4. *Selección del estadístico o función de decisión:* Su distribución en el muestreo ha de ser conocida si la hipótesis nula es verdadera. Habitualmente, el estadístico *T* será una función del estimador natural asociado al parámetro que intervenga en las hipótesis.

5. *Determinación de la región crítica:* Conocida la distribución del estadístico bajo la hipótesis nula, se halla su valor crítico en el punto a (es decir se halla k tal que $P_0(T > k) = a$). La región crítica será $T(X_1, X_2, \ldots, X_n) > k$.

6. *Cálculo del valor de la función de decisión (estadístico) para la muestra particular*: Hallamos $T_0 = T(x_1, x_2, \ldots, x_n)$.

7. *Conclusiones de tipo estadístico:* Rechazando la hipótesis nula si el valor de la función de decisión anterior cae dentro de la región crítica ($T_0 > k$), o bien aceptándola en caso contrario ($T_0 \leq k$).

8. *Interpretación práctica de resultados:* Se podrán tomar decisiones según el contraste realizado y el contexto en el que se enmarca dicho contraste.

En general, a fin de probar una hipótesis, se toma una muestra aleatoria de la población observada, se calcula un estadístico de prueba apropiado, y después se rechaza o no la hipótesis nula H_0. El conjunto de va-

lores de la estadística de prueba que lleva al rechazo de H_0 se llama región crítica, o región de rechazo de la prueba.

Ya sabemos que pueden cometerse dos tipos de errores al probar una hipótesis. Si se rechaza la hipótesis nula cuando ésta es verdadera, se cometerá un error tipo I. Si no se rechaza la hipótesis nula cuando es falsa, entonces se cometerá un error tipo II. Las probabilidades de que ocurran estos dos tipos de errores se denotan como:

$$\alpha = P\{\text{error tipo I}\} = P\{\text{rechazar } H_0 \mid H_0 \text{ es verdadera}\} = \text{nivel del test}$$

$$\beta = P\{\text{error tipo II}\} = P\{\text{no rechazar } H_0 \mid H_0 \text{ es es falsa}\}$$

Algunas veces, es conveniente trabajar con la *potencia* del contraste, que se define como:

$$\pi = 1 - \beta = 1 - P\{\text{no rechazar } H_0 \mid H_0 \text{ es es falsa}\} = P\{\text{rechazar } H_0 \mid H_0 \text{ es falsa}\}$$

Tenemos que la potencia del contraste es la probabilidad de rechazar correctamente H_0. El procedimiento general para realizar una prueba de hipótesis es especificar un valor de la probabilidad de cometer un error α de tipo I, y después diseñar un procedimiento de prueba para obtener un pequeño valor de la probabilidad de cometer el error β de tipo II.

Por regla general, se elige directamente el riesgo α y se intenta minimizar el riesgo β. El riesgo β es por lo general una función del tamaño muestral, y se controla indirectamente. Cuanto más grande sea el tamaño muestral para la prueba, tanto menor será el riesgo β.

El concepto de *p*-valor

Cuando se interpretan resultados de un contraste, las conclusiones están basadas en una regla de decisión. Ésta se establece teniendo en cuenta el riesgo que asume el investigador de cometer un error de tipo *I*, siendo la probabilidad de este error el nivel de significación α. Otras personas que quieran utilizar los resultados del experimento pueden tener en mente, sin embargo, la decisión a tomar para un nivel de significación diferente, con lo cual será útil conocer qué tipo de decisión se puede adoptar según el nivel de significación real de un contraste, basándose en los datos observados. Este concepto actuará como contrapuesto al nivel de significación elegido antes de realizar el contraste.

Se denomina *p*-valor, o nivel de significación observado o valor *p* (donde *p* indica probabilidad), al valor de α más pequeño que haga que la mues-

tra observada nos indique que se debe rechazar H_0. De este modo, las personas que vean los resultados de un experimento pueden decidir por sí mismas si el riesgo de cometer un error de tipo I es satisfactorio.

Otra forma más intuitiva de entender el concepto de *p*-valor sería la siguiente. Cuando queremos contrastar una hipótesis extraemos una única muestra de la población y obtenemos un resultado de aceptación o rechazo de la hipótesis planteada. Pero… ¿Qué habría ocurrido si hubiésemos tomado una muestra distinta de la misma población? Por ejemplo un *p*-valor igual a 0,03 contestaría a esta pregunta diciéndonos que si hubiésemos tomado 100 muestras probablemente en 97 alcanzaríamos el mismo resultado mientras que en 3 el resultado sería distinto al de la hipótesis planteada.

Muchos informes de contrastes, especialmente en las ciencias sociales, no proporcionan un valor preseleccionado de a, informándose en su lugar sobre el *p*-valor obtenido en el experimento, de manera que la persona que analiza el informe pueda tomar una decisión basándose en su propia selección de a. Por tanto, los *p*-valores representan una forma diferente de informar sobre el resultado de un contraste estadístico. Incluso si previamente se ha seleccionado un valor de a, un buen informe se acompaña, habitualmente, del *p*-valor.

Los programas informáticos relativos a contrastes de hipótesis, incluido Excel, suelen presentar el *p*-valor como una parte de los resultados mostrados en las pantallas de salida de los procedimientos donde se incorporan contrastes de hipótesis, de manera que el usuario, al elegir un nivel de significación a, puede tomar la decisión de aceptar H_0 si $P \geq a$, y rechazar H_0 si $P < a$. Por tanto, el p-valor es el valor a más pequeño que hace rechazar H_0 según la muestra observada.

El valor *P* de un contraste depende de la localización de la región crítica o de rechazo. Sea *T* el estadístico del contraste y *t* el valor tomado por dicho estadístico para los datos recogidos. Si la región crítica es bilateral, el valor *P* del contraste o nivel de significación observado, es $P(T > t) + P(T < -t) = 2P(|T| > t)$. Si la región crítica es unilateral, el valor *P* del contraste es $P(T > t)$, o bien $P(T < t)$, dependiendo de si la región crítica está en la cola de la derecha o de la izquierda.

10.2. CONTRASTES PARAMÉTRICOS

Es habitual realizar contrastes bilaterales o unilaterales sobre los valores de los parámetros de poblaciones o de sus distribuciones basados en estadísticos de distribución conocida que permitirán calcular regiones críticas a un nivel a dado.

Lo más común es realizar contrastes en poblaciones normales o con distribuciones relacionadas con la normal como la binomial. Poisson, etcétera.

También es habitual realizar contrastes paramétricos comparando las características de dos poblaciones diferentes a las que exigiremos condiciones previas de normalidad y varianzas homogéneas para que el test paramétrico pueda ser aplicado de forma robuesta.

Contrastes para poblaciones normales

Los dos cuadros siguientes resumen los contrastes más comunes para poblaciones normales. El primero de ellos presenta contrastes para el caso de parámetros de una única población normal. El segundo refleja contrastes para parámetros de dos poblaciones, generalmente utilizados para la comparación de ambas.

Hipótesis	Estadístico	Regiones críticas
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ $H_1: \mu < \mu_0$ $H_1: \mu > \mu_0$ (σ conocida)	$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \rightarrow N(0,1)$	$\lvert Z_0 \rvert > Z_{\alpha/2}$ $Z_0 < -Z_\alpha$ $Z_0 > Z_\alpha$
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ $H_1: \mu < \mu_0$ $H_1: \mu > \mu_0$ (σ desconocida)	$t_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \rightarrow t_{n-1}$	$\lvert t_0 \rvert > t_{\alpha/2,n-1}$ $t_0 < -t_{\alpha/2,n-1}$ $t_0 > t_{\alpha/2,n-1}$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$\frac{(n-1)S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$	$\chi^2_0 > \chi^2_{\alpha/2,n-1}$ ó $\chi^2_0 < \chi^2_{1-\alpha/2,n-1}$ $\chi^2_0 < \chi^2_{1-\alpha,n-1}$ $\chi^2_0 > \chi^2_{\alpha,n-1}$

Hipótesis	*Estadístico*	*Regiones críticas*
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $H_1: \mu_1 < \mu_2$ $H_1: \mu_1 > \mu_2$ (σ_1 y σ_2 dadas)	$Z_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$ $Z_0 < -Z_\alpha$ $Z_0 > Z_\alpha$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $H_1: \mu_1 < \mu_2$ $H_1: \mu_1 > \mu_2$ ($\sigma_1 \neq \sigma_2$ no dadas)	$t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ $f = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1+1} + \dfrac{(S_2^2/n_2)^2}{n_2+1}} - 2$	$\lvert t_0 \rvert > t_{\alpha/2,f}$ $t_0 < -t_{\alpha,f}$ $t_0 > t_{\alpha,f}$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $H_1: \mu_1 < \mu_2$ $H_1: \mu_1 > \mu_2$ ($\sigma_1 = \sigma_2$ no dadas)	$t_0 = \dfrac{(\bar{x}_1 - \bar{x}_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ $g = n_1 + n_2 - 2$	$\lvert t_0 \rvert > t_{\alpha/2,g}$ $t_0 < -t_{\alpha,g}$ $t_0 > t_{\alpha,g}$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$F_0 = \dfrac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha/2,n_1-1,n_2-1}$ $F_0 < F_{1-\alpha/2,n_1-1,n_2-1}$ $F_0 < -F_{1-\alpha,n_1-1,n_2-1}$ $F_0 > F_{\alpha,n_1-1,n_2-1}$

Comparación de dos poblaciones normales con datos apareados

Supongamos que tenemos dos poblaciones que siguen distribuciones normales, $N(m_1, s_1)$ y $N(m_2, s_2)$, siendo X e Y dos muestras de ambas poblaciones no necesariamente independientes, ambas de tamaño n. Un ejemplo de datos apareados sería el de una población de la cuál se extrae una muestra antes y después de la política de familia. Se quiere contrastar la igualdad de medias de las dos poblaciones basándose en las muestras, para lo que se consideran los pares de valores (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) denominados datos apareados. Sea $d_i = x_i - y_i$ $i = 1,2,...,n$, de modo que los d_i puedan considerarse como una muestra aleatoria simple de la población normal de las diferencias.

Supongamos que estamos trabajando con *muestras grandes*, y que queremos realizar el *contraste bilateral* de igualdad de medias siguiente:

Hipótesis nula $H_0: m_1 = m_2$ o $m_1 - m_2 = 0$

Hipótesis alternativa $H_1: m_1 \neq m_2$

Bajo la hipótesis nula, el estadístico $Z = \dfrac{\bar{d}}{S_d/\sqrt{n}}$ es una $N(0,1)$.

$$\bar{d} = \Sigma d_i / n \text{ y } S_d^2 = \Sigma (d_i - \bar{d})^2 / (n-1)$$

La región de aceptación del contraste viene definida mediante el intervalo dado por $(-Z_{-a/2}, Z_{a/2})$, aceptándose la hipótesis nula si se cumple:

$$\frac{|\bar{d}|}{S_d/\sqrt{n}} \leq Z_{a/2}$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula $H_0: m_1 \leq m_2$

Hipótesis alternativa $H_1: m_1 > m_2$

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-\infty, Z_a)$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{S_d/\sqrt{n}} \leq Z_a$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula H_0: $m_1 \geq m_2$

Hipótesis alternativa H_1: $m_1 < m_2$

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-Z_a, \infty)$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{S_d/\sqrt{n}} \geq Z_a$$

Para muestras grandes, un intervalo de confianza para la diferencia de medias $m_1 - m_2$ viene dado por $\bar{d} \pm Z_{a/2} S_d/\sqrt{n}$.

Supongamos que estamos trabajando con *muestras pequeñas*, y que queremos realizar el *contraste bilateral* siguiente:

Hipótesis nula H_0: $m_1 = m_2$ o $m_1 - m_2 = 0$

Hipótesis alternativa H_1: $m_1 \quad m_2$

Bajo la hipótesis nula, el estadístico $T = \frac{\bar{d}}{S_d/\sqrt{n}}$

es una *T* de Student con $n - 1$ grados de libertad.

La región de aceptación del contraste viene definida mediante el intervalo dado por $(-t_{a/2,n-1}, t_{a/2,n-1})$, aceptándose la hipótesis nula si se cumple:

$$\frac{|\bar{d}|}{S_d/\sqrt{n}} \leq t_{a/2,n-1}$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula H_0: $m_1 \pounds m_2$

Hipótesis alternativa H_1: $m_1 > m_2$

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-\infty, t_{a,n-1})$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{S_d/\sqrt{n}} \leq t_{\alpha,n-1}$$

También podemos considerar el *contraste unilateral* siguiente:

Hipótesis nula H_0: $m_1 \geq m_2$

Hipótesis alternativa H_1: $m_1 < m_2$

En este caso, la región de aceptación del contraste viene definida por el intervalo $(-t_{a,n-1}, \infty)$, aceptándose la hipótesis nula si se cumple:

$$\frac{\bar{d}}{S_d/\sqrt{n}} \geq -t_{\alpha,n-1}$$

Para muestras pequeñas, un intervalo de confianza para la diferencia de medias $m_1 - m_2$ viene dado por $\bar{d} \pm t_{\alpha/2,n-1} S_d/\sqrt{n}$.

Comparación de dos poblaciones normales independientes

Supongamos ahora que tenemos dos poblaciones independientes que siguen distribuciones normales, $N(m_1, s_1)$ y $N(m_2, s_2)$, y de tamaño igual o distinto. Un ejemplo de datos independientes sería el de dos ayuntamientos idénticos en sus variables económicas y sociales cuyas Concejalías de Familia y Asuntos Sociales aplican dos políticas incentivadoras de la natalidad distintas. Se quiere contrastar la igualdad de medias de las dos poblaciones basándose en las muestras.

Para una interpretación adecuada del test *T* de Student en muestras independientes se requiere de forma previa saber si las varianzas de las muestras pueden o no ser consideradas iguales. Homocedasticidad o igualdad de varianzas significa que las varianzas poblacionales pueden ser consideradas iguales. Este supuesto es de especial importancia en el caso de que los tamaños muestrales de cada grupo sean muy distintos. Para realizar esta prueba existen distintos test. La función Herramientas → Análisis

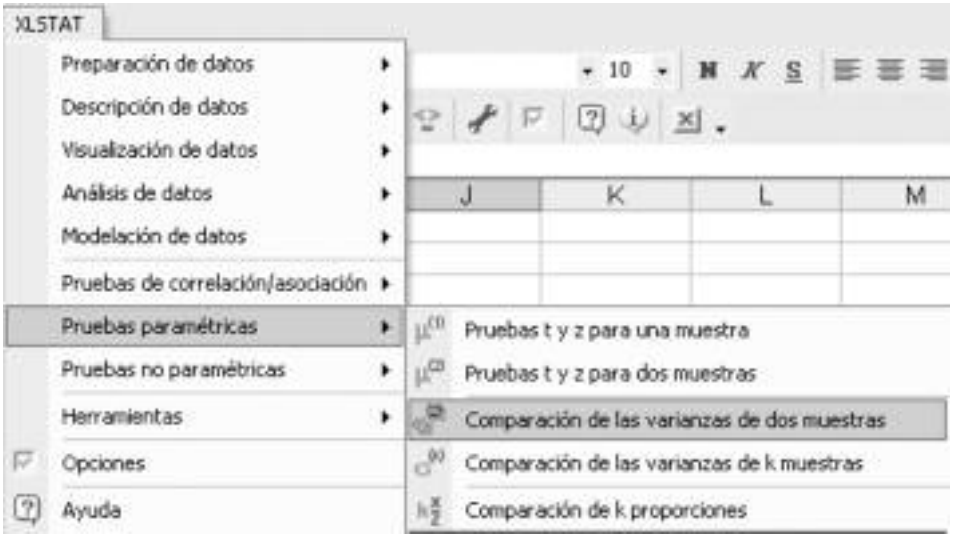


FIGURA 10-1

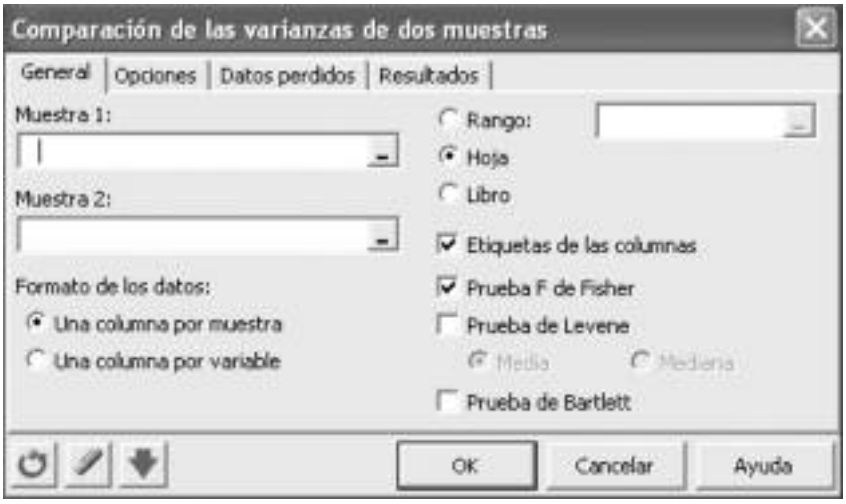


FIGURA 10-2

de Datos → Prueba F para varianzas de dos muestras permite realizar esta comparación. XLSTAT también permite hacer este cálculo (Figura 10-1) utilizando funciones adicionales a la prueba F como son el test de Levene y el test de Barlett (Figura 10-2).

Prueba F de Fisher

Sea R el Ratio entre las varianzas S_1^2 y S_2^2 correspondientes a la muestra de la población 1 y 2 respectivamente. La prueba F de Fisher se define como:

$$F = \frac{S_1^2}{RS_2^2}$$

Este estadístico sigue una distribución de *F* de Fisher con n1-1 y n2-2 grados de libertad si ambas muestras proceden de una distribución normal. Se pueden emplear tres tipos de test dependiendo de las hipótesis alternativas a elegir:

Para un test de dos colas la hipótesis nula y alternativa tomarían la siguiente forma.

$$H_0 : \sigma_1^2 = \sigma_2^2 R$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2 R$$

Si tomamos la cola de la izquierda las hipótesis que corresponden son:

$$H_0 : \sigma_1^2 = \sigma_2^2 R$$

$$H_1 : \sigma_1^2 < \sigma_2^2 R$$

Si tomamos la cola de la derecha tendremos:

$$H_0 : \sigma_1^2 = \sigma_2^2 R$$

$$H_1 : \sigma_1^2 > \sigma_2^2 R$$

Test de Levene

El test de Levene es ampliamente utilizado en estadística para comparar la igualdad entre dos o más varianzas. Es un test de dos colas similar al visto con anterioridad aunque más complejo. El lector interesado puede acudir a http://en.wikipedia.org/wiki/Levene's_test

El test de Levene sigue una distribución F con 1 y n1+n2-2 grados de libertad.

Test de homogeneidad de varianzas de Bartlett

Test de dos colas análogo a los anteriores pero sensible a la normalidad de los datos que en caso de incumplirse o de estar en duda se debería recurrir al test de Levene. Sin embargo este test, en caso de cumplirse la normalidad en los datos es más potente que los anteriores. Este test se distribuye según una χ^2 con un grado de libertad.

Test T y Z en muestras independientes

Sean dos muestras con varianzas S_1^2 y S_1^2 de tamaños n_1 y n_2 y con medias μ_1 y μ_2 respectivamente. Sea D la diferencia asumida en nuestra hipótesis nula (por ejemplo $D = 0$ si asumimos que las dos medias son iguales).

Una vez comprobado si las varianzas son o no iguales en los grupos podemos utilizar los test T de Student y Z. Si consideramos que las dos muestras tienen la misma varianza entonces la varianza común se estima como:

$$S^2 = \frac{\left[(n_1-1)S_1^2+(n_2-1)S_2^2\right]}{(n_1+n_2-2)}$$

Entonces el test T sigue una distribución de Sudent con n1 + n2 - 2 grados de libertad definido por la siguiente expresión:

$$t = \frac{\left[(\mu_1-\mu_2-D)\right]}{S\sqrt{1/n_1+1/n_2}}$$

si consideramos que las varianzas son diferentes el test toma la siguiente forma

$$t = \frac{\left[(\mu_1-\mu_2-D)\right]}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$$

donde el número de grados de libertad se calcula a través de la propuesta de Salterwhaite

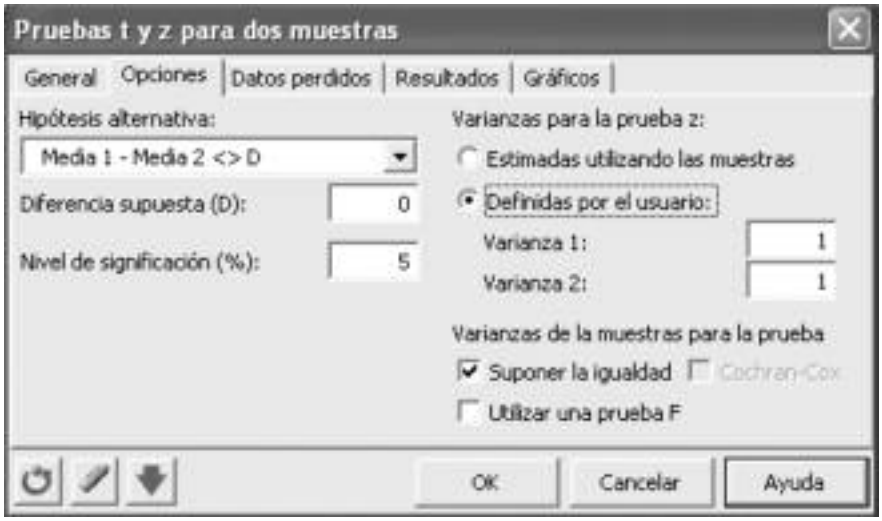


FIGURA 10-3

$$df = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$

donde con n1 = n2 los grados de libertad serían df = 2(n-1)

En el caso del test Z se asume que la varianza de la población es conocida por lo que el usuario puede introducir su valor en la pestaña de opciones de la herramienta de XLSTAT (Figura 10-3).

El test Z toma la siguiente forma que se distribuye como una normal y toma la siguiente forma:

$$z = \frac{\left[\left(\mu_1 - \mu_2 - D\right)\right]}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Contrastes para los parámetros de las variables Binomial y Poisson

Estos contrastes se basan en los resultados que se enuncian en los párrafos siguientes.

Sea $p = x/n$ la proporción poblacional estimada del número de veces que aparece un suceso de Bernouilli de entre n repeticiones de un experimento (x es el número de veces que aparece el suceso). Si $np(1- p) > 25$, el estadístico:

$$Z=\frac{X+1/2-np}{\sqrt{np(1-p)}}$$

sigue una distribución N (0,1), siendo X el número de veces que aparece en la muestra el suceso considerado.

Cuando la población sigue una distribución de Poisson de parámetro 1con $n1> 25$, el estadístico;

$$Z=\frac{X+1/2-n\lambda}{\sqrt{n\lambda}}$$

sigue una distribución normal (0,1).

Hipótesis	Estadístico	Regiones críticas
$H_0: p = p_0$ $H_1: p \neq p_0$ $H_1: p < p_0$ $H_1: p > p_0$ (Binomial)	$Z_0 = \frac{X + \frac{1}{2} - np_0}{\sqrt{np_0(1-p_0)}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$ $Z_0 < -Z_\alpha$ $Z_0 > Z_\alpha$
$H_0: \lambda = \lambda_0$ $H_1: \lambda \neq \lambda_0$ $H_1: \lambda < \lambda_0$ $H_1: \lambda > \lambda_0$ (Poisson)	$Z_0 = \frac{X + \frac{1}{2} - n\lambda_0}{\sqrt{n\lambda_0}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$ $Z_0 < -Z_\alpha$ $Z_0 > Z_\alpha$
$H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$ $H_1: p_1 < p_2$ $H_1: p_1 > p_2$ (Diferencia Binomiales)	$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$ $Z_0 < -Z_\alpha$ $Z_0 > Z_\alpha$

El tratamiento de la información faltante constituye una de las tareas previas a cualquier análisis. Cuando se aplica un método de análisis multivariante sobre los datos disponibles puede ser que no exista información para determinadas observaciones y variables. Estamos entonces ante valores ausentes o valores *missing*. La presencia de esta información faltante puede deberse a un registro defectuoso de la información, a la ausencia natural de la información buscada o a una falta de respuesta (total o parcial).

Contrastes del coeficiente de correlación y regresión

Es posible contrastar la hipótesis de que el coeficiente de correlación poblacional sea cero o un valor fijo distinto de cero.

Bajo la hipótesis nula de que el coeficiente de correlación poblacional sea nulo ($r = 0$), el estadístico:

$$T=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

se distribuye según una *t* de Student con *n*-2 grados de libertad, siendo *r* el coeficiente de coeficiente de correlación muestral y *n* el tamaño de la muestra.

Este estadístico permite *contrastar la hipótesis nula de que el coeficiente de correlación poblacional es cero*. Para ello, se halla el valor *k* tal que $P(T_{n-2} \geq k) = a$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico T para los datos dados de la muestra es mayor que *k* se rechaza la hipótesis nula de incorrelación al nivel fijado α. En caso contrario se acepta la incorrelación.

Bajo la hipótesis nula de que el coeficiente de correlación poblacional sea $r_0 \neq 0$, se tiene:

$$Z=\frac{1}{2}Ln\frac{1+r}{1-r}\rightarrow N\left(\frac{1}{2}Ln\frac{1+\rho_0}{1-\rho_0}+\frac{\rho_0}{2(n-1)},\sqrt{\frac{1}{n-3}}\right)$$

Este estadístico permite *contrastar la hipótesis nula de que el coeficiente de correlación poblacional es un valor dado* ($r=r_0$). Para ello, se halla el valor *k* tal que $P(Z \geq k) = a$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico Z para los datos dados de la muestra es mayor que *k* se rechaza la hipótesis nula de que el coeficiente de correlación poblacional r vale r_0. En caso contrario se acepta $r=r_0$.

Asimismo, también es posible contrastar la hipótesis de que le coeficiente de regresión poblacional de una variable sobre otra sea un valor fijo.

Bajo la hipótesis nula de que el coeficiente de regresión poblacional de *Y* sobre *X* sea b_{12}, el estadístico:

$$T=\frac{S_2\sqrt{n-2}}{S_1\sqrt{1-r^2}}(b_{12}-\beta_{12})$$

se distribuye según una *t* de Student con *n*-2 grados de libertad, siendo b_{12} el coeficiente de regresión muestral de *Y* sobre *X, r* el coeficiente de correlación muestral de *X* e *Y, S*$_1$ y *S*$_2$ las cuasivarianzas muestrales de *X* e *Y*, y *n* el tamaño de la muestra. Este estadístico permite *contrastar la hipótesis nula de que el coeficiente de regresión poblacional de Y sobre X es* b_{12}. Para ello, se halla el valor *k* tal que $P(T_{n-2} \geq k) = a$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico T para

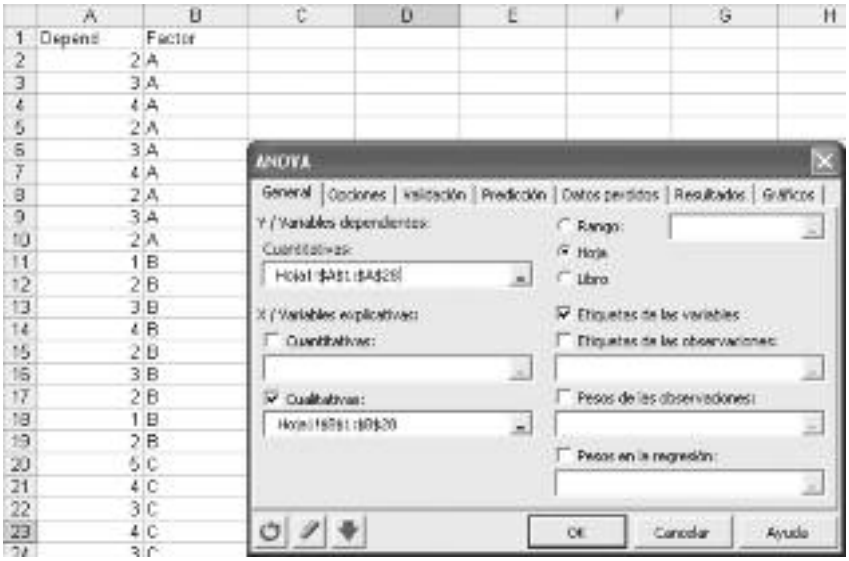


FIGURA 10-4

los datos dados de la muestra es mayor que k se rechaza la hipótesis nula de que el coeficiente de regresión poblacional de Y sobre X sea b_{12} al nivel fijado α. En caso contrario se acepta b_{12} como coeficiente de regresión de Y sobre X.

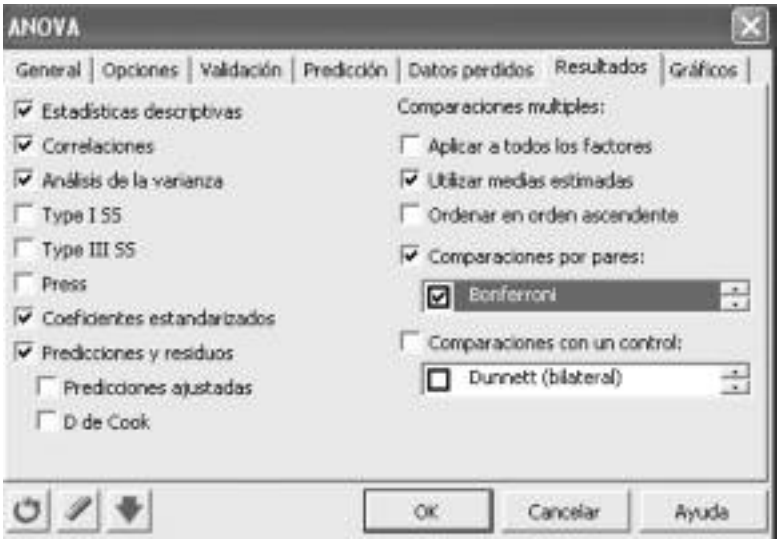


FIGURA 10-5

Análisis de Varianza

Hasta ahora todos los contrastes que se han visto comparaban dos muestras que diferían entre sí en dos categorías de un factor. Sin embargo existen factores con más de dos categorías. Por ejemplo tipos de familia: nucleares, monoparentales, otras; origen del ciudadano: nativo, no nativo, primera generación; por el número de hijos: familias sin hijos, con hijos no numerosas, familias numerosas, etc. El análisis de varianza amplía el test paramétrico de la T a estos casos. Llamaremos variable dependiente a la variable continua cuya media queremos contrastar y variable independiente o factor a la variable cuyas categorías nominales u ordinales nos sirven para establecer los grupos. Para la correcta aplicación de esta prueba es necesario contrastar de forma previa la normalidad de la distribución de la variable dependiente en cada grupo (XLSTAT → Descripción de Datos → Pruebas de Normalidad) y la homogeneidad de varianzas entre los grupos utilizando el test de Levene.

En Excel este procedimiento puede aplicarse a través de Herramientas → Análisis de Datos → Análisis de Varianza de un Factor. Sin embar-

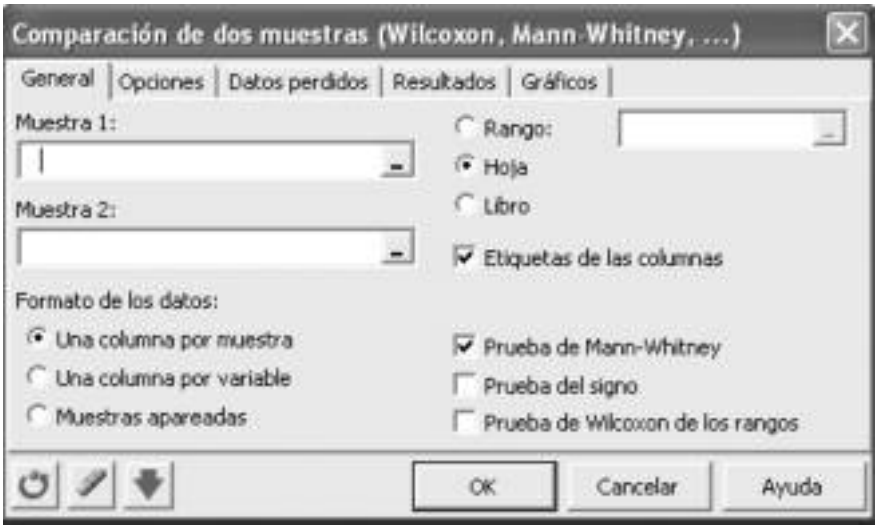


FIGURA 10-6

go esta prueba no permite contrastar la homogeneidad de varianzas ni realizar test de diferencias de medias dos a dos entre todas las categorías del factor. La herramienta XLSTAT → Modelación de Datos → ANOVA ofrece más opciones para un análisis más adecuado (Figura 10.4).

En el análisis de varianza resulta muy interesante contrastar las diferencias entre cada par de muestras dentro del factor además de aceptar o rechazar la hipótesis de que todas las medias son iguales. Para ello en la pestaña Resultados existe la opción comparación por pares que permite indagar acerca de estas diferencias (Figura 10-5). Dentro de éstos el más aplicado en la práctica empírica es el test de Bonferroni basado en la prueba T. Este test tiene en cuenta que a medida que aumenta el número de comparaciones dos a dos también aumenta la probabilidad de encontrar por azar una relación significativa (aunque puediera ser espúrea) y por ello corrige el valor de *a.*

$$\alpha' = \frac{\alpha}{g(g-1)/2}$$

10.3. CONTRASTES NO PARAMÉTRICOS

En los contrastes de hipótesis realizados hasta ahora hemos supuesto que la población de la cual tomábamos las muestras pertenecía a una determinada familia de distribuciones (normalidad, varianza homogéneas) y, por tanto, la estimación o contraste de la hipótesis previamente definida nos permitía especificar totalmente esa población.

A continuación vamos a analizar algunos procedimientos que no exigen ningún supuesto, o muy pocos, acerca de la familia de distribuciones a que pertenece la población, procedimientos que, además, soportan observaciones en donde las mediciones se realizan en forma cualitativa (clasificando caracteres cualitativos) o bien se refieren a alguna característica ordenable como escalas de Likert. A tales procedimientos se les denomina *contrastes no paramétricos.*

Nuevamente distinguiremos para diferencia de medias entre dos grupos apareados e independientes. Para realizar este contraste acudiremos a la utilidad XLSTAT → Pruebas no paramétricas → Comparación de los muestras (Figura 10-6).

Comparación de dos muestras

Para dos muestras de tamaño n1 y n2 independientes usaremos el test de Mann-Whitney. Si juntamos las n1 + n2 = n observaciones y las ordenamos a partir de la variable dependiente Y podemos asignar rangos R a las n observaciones (1 a la más pequeña y n a la más grande) resolviendo los empates dando un rango promedio. A partir de aquí podemos definir el estadístico S1 como la suma de los rangos asignados a la muestra 1 y S2 como la suma de los rangos asignados a la muestra 2. El estadístico U adopta la siguiente forma en cada grupo:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - S_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - S_2$$

Asumimos que U_1 y U_2 deben ser aproximadamente iguales ya que las dos poblaciones provienen de poblaciones idénticas. Si son muy distintos existirá evidencia de que las dos muestras proceden de poblaciones distintas y por tanto rechazaremos la hipótesis de que los promedios poblacionales sean iguales.

Para el caso de dos muestras apareadas donde X es el vector de observaciones iniciales de una muestra n e Y es el vector de observaciones finales también de tamaño n podemos definir n pares de valores (x_i, y_i) de los cuales calculamos sus diferencias en valor absoluto.

$$D_i = |X_i - Y_i|$$

Sin considerar las diferencias nulas asignaremos rangos R desde 1 hasta m a las diferencias no nulas asignando 1 a la diferencia más pequeña y m ≤ n a la menor. A continuación se suman por un lado los rangos positivos Xi > Yi y llamemos S+ a esta suma y por otro lado se suman los rangos negativos Xi<Yi y llamemos S- a esta otra suma. Si suponemos que las dos muestras provienen de ambas poblaciones entonces la mediana de X debe coincidir con la mediana de Y, de esta afirmación es fácil deducir que:

$$S_+ = \sum R_i^+ \approx S_- = \sum R_i^-$$

Por lo que una fuerte diferencia entre ambas expresiones hará dudar de que las dos muestras procedan de la misma población o de dos poblaciones idénticas. Finalmente, en el caso de dos muestras apareadas utilizamos la prueba de Wilcoxon también presente en XLSTAT (Figura 10-16).

Comparación de más de dos muestras independientes

El equivalente no paramétrico del análisis de varianza es la prueba de Kruskal-Wallis. Por tanto este test es robusto ante muestras que no cumplan los supuestos de normalidad y homocedasticidad además de permitir trabajar con datos ordinales. De forma análoga al test de Mann-Whitney consideraremos J muestras aleatorias e independientes de tamaños n1, n2, …, nj extraídas de la misma población o de poblaciones idénticas donde n = n1 + n2 + …+nj. Asignaremos rangos desde 1 a n a ese conjunto de observaciones como si se tratara de una sola muestra y en caso de empa-

tes se asigna el promedio de los rangos empatados. Sea Rij los rangos asignados a las observaciones i de la muestra j. Llamaremos Rj a la suma de los rangos asignados a las nj observaciones de la muestra j. Tendremos:

$$R_j = \sum_{i}^{n_j} R_{ij}$$

$$\bar{R}_j = \frac{R_j}{n_j}$$

De esta manera si las J poblaciones son idénticas los Rj de las distintas muestras serán parecidas. El estadístico para este cálculo es:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{J} \frac{R_j^2}{n_j} - 3(n+1)$$

Bajo la hipótesis nula de que los J promedios poblacionales son iguales, el estadístico H se distribuye según una Chi-cuadrado con J-1 grados de libertad.

10.4. TABLAS DE CONTINGENCIA

Además de las diferencias de medias analizadas, en la investigación social de la familia resulta interesante comprobar si las categorías de dos variables aparecen asociadas. Por ejemplo podría ser interesante estudiar si el nivel de estudios (alto, medio, bajo) de los hijos tiene alguna relación con el nivel de estudios de sus padres (alto, medio, bajo). Podemos también estudiar si la valoración (muy buena, buena, normal, mala, muy mala) que los individuos tienen de las iniciativas de políticas de familia del ayuntamiento está asociada a que éstos disfruten de este tipo de ayudas y de sus cuantías (0 euros, menores de 500 euros, entre 501 y 1.000 euros y de más de 1.000 euros). A este análisis donde se cruzan las categorías de dos variables se le denomina análisis de tablas de contingencia.

Si *X* e *Y* son dos variables observadas, la distribución bidimensional (X,Y) será (x_i, y_j, n_{ij}). Cada frecuencia corresponde ahora a un par de valores (variables cuantitativas) o modalidades (variables cualitativas): el primer elemento del par corresponde al valor de la primera característica observada, mientras que el segundo hace referencia a la segunda de tales características, y el tercero a la frecuencia conjunta. Evidentemente, sería posible realizar un estudio por separado de la distribución de *X* e *Y*, y resumir estos caracteres por medio de sus medidas de posición

y dispersión descritas en el capítulo anterior; tales distribuciones recibirán el nombre de *distribuciones marginales*. Sin embargo, nuestro interés en este punto se centra en el análisis simultáneo de ambas características; es decir, en la *distribución conjunta* de las mismas, con el fin de establecer si existe relación entre ellas y en qué grado. Los pares que contienen los valores de las variables o atributos junto con sus correspondientes frecuencias, suelen disponerse en una tabla de doble entrada, que recibe el nombre de *tabla de correlación* en el caso de que ambos caracteres sean cuantitativos, y *tabla de contingencia* cuando son cualitativos. Estos dos tipos de tablas serán objeto de nuestra atención en los apartados siguientes.

Queremos estudiar conjuntamente dos variables cuantitativas o cualitativas X e Y, sobre una población, apareciendo X con h niveles e Y con k. Para ello seleccionamos una muestra de tamaño N y la sometemos a observación, disponiendo los resultados en una tabla de doble entrada, donde $x_1, \ldots, x_h$ e $y_1, \ldots, y_k$ representan los valores observados para cada variable, y n_{ij} la frecuencia absoluta conjunta, es decir, las veces que aparecen simultáneamente el valor i-ésimo de X y j-ésimo de Y.

$X, Y \rightarrow$ $\downarrow$	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_k	$n_{i\cdot}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_h	n_{h1}	n_{h2}	$\cdots$	n_{hj}	$\cdots$	n_{hk}	$n_{h\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	N

n_{ij} = frecuencia absoluta del valor (X_i, Y_j) de la distribución conjunta (X,Y).

$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$ = frecuencia absoluta del valor X_i de la variable marginal X.

$n_{\cdot j} = \sum_{i=1}^{h} n_{ij}$ = frecuencia absoluta del valor Y_j de la variable marginal Y.

$f_{ij} = \frac{n_{ij}}{N}$ = frecuencia relativa del valor (X_i, Y_j) de la distribución conjunta (X,Y).

$f_{i\cdot} = \frac{n_{i\cdot}}{N}$ = frecuencia relativa del valor X_i de la variable marginal X.

$f_{\cdot j} = \frac{n_{\cdot j}}{N}$ = frecuencia relativa del valor Y_j de la variable marginal Y.

Se cumple que:

$$\sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{h} n_{\cdot j} = \sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = N \text{ y } \sum_{i=1}^{k} f_{i\cdot} = \sum_{j=1}^{h} f_{\cdot j} = \sum_{i=1}^{k}\sum_{j=1}^{h} f_{ij} = 1$$

En esta tabla, $n_{i\cdot}$ y $n_{\cdot j}$ nos proporcionan las *frecuencias marginales*. Es decir, el número de veces que aparece el valor i-ésimo de X, con independencia de cuál sea el valor de Y, es $n_{i\cdot}$, y el número de veces que aparece el valor j-ésimo de Y, independientemente de cuál sea el valor de X con el que se da conjuntamente Y, es $n_{\cdot j}$. De esta forma tenemos que las *distribuciones marginales* de X e Y vienen dadas por $(x_i; n_{i\cdot})$ y $(y_j; n_{\cdot j})$. Estas distribuciones marginales pueden expresarse como sigue:

X	$n_{i\cdot}$
x_1	$n_{1\cdot}$
x_2	$n_{2\cdot}$
⋮	⋮
x_k	$n_{k\cdot}$
	N

Y	$n_{\cdot j}$
y_1	$n_{\cdot 1}$
y_2	$n_{\cdot 2}$
⋮	⋮
y_h	$n_{\cdot h}$
	N

Dado que estas distribuciones marginales son variables unidimensionales, es posible calcular todo tipo de medidas de centralización, dispersión y forma, mediante los procedimientos ya vistos en el capítulo anterior.

A partir de la tabla de correlación es posible formar un nuevo tipo de distribuciones, que denominaremos *distribuciones condicionadas* debido a que para su obtención es preciso definir previamente una condición. Esta condición hará referencia a la fijación *a priori* de un valor (o valores) de una de las variables, para posteriormente calcular la distribución de la otra variable sujeta a esa condición. Si fijamos la variable Y en el valor y_2 (podríamos fijar más de un único valor), la distribución de la variable X condicionada a que Y tome el valor y_2 vendrá dada por:

$X/Y=y_2$	$n_{i/j=2}$
x_1	n_{12}
x_2	n_{22}
⋮	⋮
x_h	n_{h2}
	$n_{\cdot 2}$

Donde $X/Y=y_2$ nos dará los valores que puede tomar la variable X cuando Y toma el valor y_2, y $n_{i/j=2}$ nos da las frecuencias con que se presenta cada uno de los valores.

En general, dado que se pueden establecer condiciones sobre Y y X calculando posteriormente la distribución de X o Y sujeta a esa condición, nos encontramos distribuciones que, de manera genérica, tendrán la forma:

X/Y_j	$n_{i/j}$
x_1	n_{1j}
x_2	n_{2j}
⋮	⋮
x_h	n_{hj}
	$n_{\cdot j}$

Y/X_i	$n_{j/i}$
y_1	n_{i1}
y_2	n_{i2}
⋮	⋮
y_k	n_{ik}
	$n_{i\cdot}$

Dado que estas distribuciones condicionadas son variables unidimensionales, es posible calcular todo tipo de medidas de centralización, dispersión y forma, mediante los procedimientos ya vistos en el capítulo anterior.

Para todas las distribuciones condicionadas, será posible trabajar con frecuencias relativas en vez de con frecuencias absolutas. Tenemos lo siguiente:

$$f_{j/i}=\frac{f_{ij}}{f_{i\cdot}}=\frac{n_{ij}/N}{n_{i\cdot}/N}=\frac{n_{ij}}{n_{i\cdot}} \quad f_{i/j}=\frac{f_{ij}}{f_{\cdot j}}=\frac{n_{ij}/N}{n_{\cdot j}/N}=\frac{n_{ij}}{n_{\cdot j}}$$

$$\sum_{j=1}^{k} f_{j/i}=\sum_{j=1}^{k}\frac{n_{ij}}{n_{i\cdot}}=\frac{1}{n_{i\cdot}}\sum_{j=1}^{k} n_{ij}=\frac{n_{i\cdot}}{n_{i\cdot}}=1 \quad \sum_{i=1}^{h} f_{i/j}=\sum_{i=1}^{h}\frac{n_{ij}}{n_{\cdot j}}=\frac{1}{n_{\cdot j}}\sum_{i=1}^{h} n_{ij}=\frac{n_{\cdot j}}{n_{\cdot j}}=1$$

Otra relación importante entre distribuciones condicionadas, marginales y conjunta es la siguiente:

$$\frac{n_{ij}}{N}=\frac{n_{i.}}{N}\frac{n_{ij}}{n_{i.}}=\frac{n_{.j}}{N}\frac{n_{ij}}{n_{.j}}\Rightarrow f_{ij}=f_{i.}f_{ij/i}=f_{.j}f_{i/j}$$

¿Cómo podemos detectar de modo sencillo la existencia de independencia entre dos variables? ¿Qué instrumentos estadísticos son los que nos permiten señalar la ausencia de tal relación? Para detectar la no presencia de asociación entre dos caracteres analizados sobre la misma población, se procede a elaborar la tabla de correlación (para variables cuantitativas) o de contingencia (para variables cualitativas), y se calculan las respectivas distribuciones conjuntas, marginales y condicionadas. Las variables son independientes si se cumple cualquiera de las dos siguientes condiciones equivalentes:

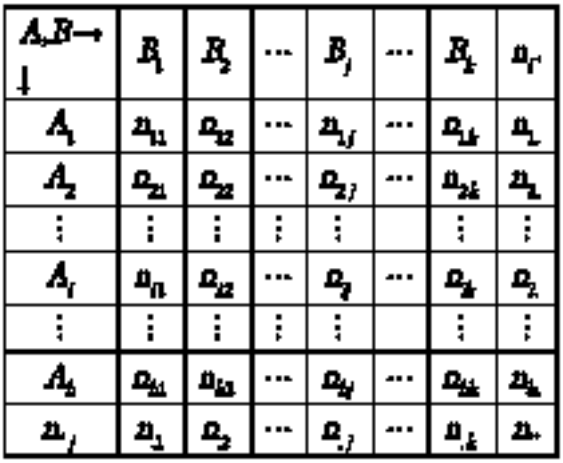
- Las frecuencias relativas condicionadas coinciden con sus respectivas frecuencias relativas marginales, lo que nos indica que el condicionamiento, en cuanto tal, no existe. Ha de cumplirse que $f_{i/j} = f_{i\cdot} = n_{i.} /N$ y $f_{j/i} = f_{\cdot j} = n_{.j} /N$ para todo *i,j*.

- La frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales, es decir, $n_{ij} /N = (n_{i.} /N)*(n_{.j} /N) \Leftrightarrow f_{ij} = f_{i\cdot} f_{\cdot j}$ para todo *i,j*.

Cuando las dos variables son independientes, la covarianza es cero, aunque debemos señalar que el recíproco no es siempre cierto (es decir, la covarianza nula no implica necesariamente que ambas variables sean independientes).

Contraste chi-cuadrado de independencia entre los atributos de dos variables en una tabla de contingencia

Se pretende contrastar, con un nivel de significación ?, la hipótesis nula de independencia entre dos variables que pueden ser dos atributos A *y* B, que presentan *h y k* niveles exhaustivos y mutuamente excluyentes, respectivamente. Para ello se toma una muestra aleatoria simple de tamaño N de una población bidimensional, y los N valores muestrales se clasifican en una tabla de doble entrada, denominada Tabla de Contingencia, en función de los niveles de *A* y de *B* que posean. Así, la información muestral obtenida se clasifica como indica la tabla siguiente:



Tenemos que la observación simultánea de dos atributos da lugar a una tabla de doble entrada, en donde n_{ij} indica el número de objetos o individuos que poseen conjuntamente las modalidades indicadas en la fila i-ésima y en la columna j-ésima de la tabla de contingencia. Sabemos que dos atributos *A* y *B* son independientes cuando entre ellos no existe ningún tipo de influencia mutua. Si dos atributos, *A y B*, son independientes estadísticamente, la frecuencia relativa conjunta será igual al producto de las frecuencias marginales respectivas. Para que *A* y *B* sean independientes habrá de cumplirse que $n_{ij} = (n_i.n_.j)/N$ para todo *i,j*. En la práctica basta con que la relación se verifique para $(h\text{-}1)(k\text{-}1)$ valores de n_{ij}, ya que entonces se verificará para todos los restantes.

Si designamos por n_{ij} la frecuencia conjunta correspondiente a las modalidades A_i del atributo *A* y B_j de *B*, y por n_{ij}' la frecuencia teórica que correspondería en el caso de que ambos atributos fuesen independientes, esto es, $n_{ij}' = (n_i.n_.j)/N$, $i=1,...,h$, $j=1,...,k$, siendo *N* el total de elementos que se estudian, definimos el *coeficiente de contingencia* c^2 como sigue:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(n_{ij}-n_{ij}')^2}{n_{ij}'}$$

Este coeficiente también se denomina en la literatura estadística *cuadrado de la contingencia*, y puede expresarse de forma más sencilla para el cálculo como sigue:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_{ij}'} - N$$

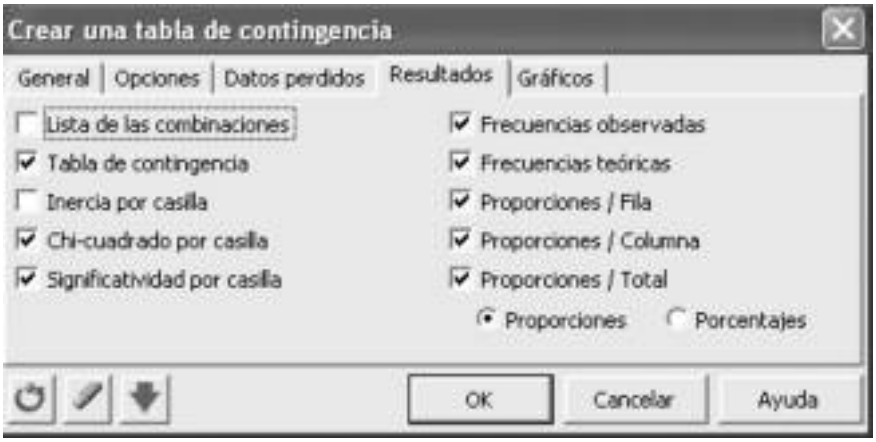


FIGURA 10-7

El coeficiente de contingencia χ^2 se utiliza para realizar un contraste formal para la hipótesis nula de independencia de los atributos *A* y *B* cuya información muestral se recoge en la tabla de contingencia dada. La hipótesis alternativa es la existencia de asociación entre los atributos *A* y *B*. El contraste se basa en que, bajo la hipótesis nula de independencia de los atributos *A* y *B*, el estadístico χ^2 se distribuye según una chi-cuadrado con $(h\text{-}1)(k\text{-}1)$ grados de libertad.

Para realizar el contraste se halla el valor *k* tal que $P(\chi^2_{(h\text{-}1)(k\text{-}1)} \geq k) = a$, siendo *a* el nivel de significación establecido para el contraste. Si el valor del estadístico χ^2 para los datos dados de la tabla de contingencia es mayor que *k* se rechaza la hipótesis nula de independencia de los atributos *A* y *B* al nivel fijado α. En caso contrario se acepta la independencia.

Cuando el tamaño muestral es pequeño (*N* menor que 150) se utiliza el test exacto de Fisher para contrastar la independencia de atributos. En este caso suele introducirse una corrección por continuidad en el estadístico de la chi-cuadrado, tomando en su lugar para el contraste de independencia el estadístico corregido de Yates, cuya expresión es la siguiente:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{\left(|a_{ij}' - a_{ij}| - \frac{1}{2}\right)^2}{a_{ij}'}$$

Como concepto contrario al de independencia tenemos el de *asociación.* Se dice que *A y B* están asociados cuando aparecen juntos en mayor

número de casos que el que cabría esperar si fuesen independientes. Según que esa tendencia a coincidir o no coincidir esté más o menos marcada, tendremos distintos grados de asociación. Para medirlos se han ideado diversos procedimientos, denominados *coeficientes de asociación*, entre los que destacan los ya estudiados en el capítulo 3.

La herramienta XLSTAT permite realizar tales cálculos a través de XLSTAT → Preparación de Datos → Crear una tabla de contingencia. Mediante esta opción (Figura 10-7) podemos solicitar información acerca de todos los estadísticos comentados con anterioridad.

Si en lugar de tener la información por columnas la tenemos ya construida en forma de tabla de contingencia entonces usaremos XLSTAT → Descripción de datos → Tabla de contingencia. En esta función se permite hacer los mismos análisis que en la utilidad anterior pero cambiando la forma de introducir los datos.

10.5. ANÁLISIS DE CORRELACIONES

Terminaremos el estudio de la asociación entre dos variables para el caso en el que ambas son continuas. Así, nos puede interesar determinar si los ingresos familiares están correlacionados con las ayudas recibidas o si un indicador de pobreza está correlacionado con la tasa de paro.

Se llama correlación al grado de dependencia mutua entre dos variables. El *coeficiente de correlación* intenta medir la intensidad con que dos variables están relacionadas. Este concepto está directamente relacionado con el concepto de curva de regresión. Mediante la regresión simple mínimo cuadrática, se expresa la estructura funcional de la relación existente entre dos variables, ajustando la nube de puntos dada por los pares de valores de las dos variables a una curva de la forma mejor posible (minimizando la varianza del error). El ajuste será de la forma $Y=f(x)+e$ o $X=f(Y)+e$, donde e denota el error cometido cuya varianza debe ser mínima. El coeficiente de correlación mide la calidad de ese ajuste.

Cuando la curva es una recta, la regresión se llama lineal, y en este caso el coeficiente de correlación se llama *coeficiente de correlación lineal,* y mide el grado de asociación lineal que existe entre las variables. El ajuste será de la forma $Y=a+bX+e$ (*recta de regresión de Y sobre X*), o $X=c+dY+e$ (*recta de regresión de X sobre Y*), donde $a =$, $b =$, $c =$ y $d =$.

A los parámetros a y b se les denomina *coeficientes de regresión de Y sobre X,* y a los parámetros c y d se les llama *coeficientes de regresión de*

X sobre Y. También se pueden expresar las rectas de regresión de *Y* sobre *X* y *X* sobre *Y* respectivamente de la forma.

Si suponemos el ajuste de la forma *Y=a+bX+e* (*recta de regresión de Y sobre X*), el criterio de mínimos cuadrados considera que la función que mejor se ajusta a los datos es la que minimiza la varianza del error *e*, lo que es equivalente a minimizar:

$$S(a,b)=\sum_{i,j} e_{ij}^2=\sum_{i,j}(y_j-(a+bx_i))^2$$

Derivando respecto de los parámetros *a* y *b* e igualando a cero tenemos:

$$\left.\begin{aligned}\frac{\partial S(a,b)}{\partial a}&=2\sum_{i,j}(y_j-(a+bx_i))(-1)=0\\ \frac{\partial S(a,b)}{\partial b}&=2\sum_{i,j}(y_j-(a+bx_i))(-x_i)=0\end{aligned}\right]\Rightarrow\left[\begin{aligned}\sum_j y_j&=an+b\sum_i x_i\\ \sum_{i,j} y_j x_i&=a\sum_i x_i+b\sum_i x_i^2\end{aligned}\right.$$

Las soluciones *a* y *b* de este *sistema de ecuaciones normales* son *a* = y *b* =, lo que hace que la recta de regresión de Y sobre X sea:

$$Y=a+bX=\ \bar{y}-\bar{x}S_{xy}/S_x^2+S_{xy}/S_x^2 x\Rightarrow y-\bar{y}=(x-\bar{x})S_{xy}/S_x^2$$

Razonando de forma similar, se obtienen la recta de regresión de *X* sobre *Y*.

La expresión del coeficiente de correlación lineal entre las variables *X* e *Y* viene dado por la expresión:

$$r=\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(x_i-\bar{x})\ (y_j-\bar{y})\ n_{ij}}{\sqrt{\sum_{i=1}^{k}(x_i-\bar{x})^2\sum_{j=1}^{k}(y_j-\bar{y})^2}}$$

Si *r = 1* existe correlación perfecta positiva, y la relación funcional entre ambas variables es exacta y positiva, variando las dos en el mismo sentido (al aumentar una aumenta la otra, y al disminuir una disminuye la otra). Si *r = - 1* existe correlación perfecta negativa, y la relación funcional entre ambas variables es exacta y negativa, variando las dos en el sentido opuesto (al aumentar una disminuye la otra, y al disminuir una aumenta la otra). Si *r = 0* la correlación es nula, y las varia-

bles no están asociadas, siendo imposible encontrar una relación funcional entre ellas.

Si $0<r<1$ la correlación es positiva, pero el grado de asociación entre las dos variables será mayor a medida que r se acerca más a 1, y será menor a medida que r se acerca más a cero. Si $-1<r<0$ la correlación es negativa, pero el grado de asociación entre las dos variables será mayor a medida que r se acerca más a -1, y será menor a medida que r se acerca más a cero.

El cuadrado del coeficiente de correlación r^2, denotado en general por R^2, se denomina *coeficiente de determinación* y representa el porcentaje de variabilidad de la variable dependiente que es explicada por la regresión. Dada su definición, R^2 puede expresarse de forma general en función de la varianza de Y y de la varianza residual como sigue:

$$R^2 = \frac{\text{Varianza de la regresión}}{\text{Varianza de } Y} = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} = 1 - \frac{\sigma_e^2}{\sigma_y^2} = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(y_i - (a - b\overline{x_i}))^2 n_{ij}}{\sigma_y^2}$$

En general, se trata de una medida de la bondad del ajuste por regresión. Si R^2 se aproxima a la unidad el ajuste es bueno y si R^2 se acerca a cero el ajuste es malo. Esta definición e interpretación de R^2 es válida para cualquier tipo de ajuste aunque no sea lineal.

En cuanto a relación entre correlación e independencia, se observa que al definir el coeficiente de correlación lineal como $r = S_{xy}/(S_x S_y)$, si las variables son independientes estarán incorrelacionadas, ya que $r=0$ debido a que S_{xy} es cero cuando hay independencia. Ahora bien, el recíproco no es necesariamente cierto, ya que dos variables pueden estar incorrelacionadas linealmente y ser dependientes, puesto que al ser $r=0$, lo único que podemos decir es que la asociación lineal es nula, pero esas variables pueden depender según otro tipo de asociación (parabólica, exponencial, etc.).

Coeficiente de correlación por rangos

A veces nos interesa ver si la opinión, preferencias o grado de uso de instalaciones públicas están correlacionadas. Sin embargo las medidas ordinales no cumplen con el supuesto de que sean variables cuantitativas. Es por tanto muy típico considerar, sobre todo en variables cualitativas, el coeficiente de correlación entre los rangos de los valores de las variables. Se entiende por rango de un valor de una variable el lugar que ocupa dicho valor en el conjunto total de valores de la variable, suponiendo una ordenación de

menor a mayor. Sean A_i y B_i las diferentes modalidades de dos variables cualitativas X e Y. Sean x_i e y_i los rangos o números de orden que le corresponden a A_i y B_i, supuestas ordenadas estas modalidades, con la escala que se determine, y de menor a mayor. Se define el coeficiente de correlación por rangos de Spearman para las variables cualitativas X e Y como el coeficiente de correlación lineal de las variables cuyos valores son x_i e y_i.

Este valor se utiliza para medir el grado de asociación de las variables cualitativas X e Y basándonos en la concordancia o discordancia de las clasificaciones por rangos de sus modalidades. El coeficiente de correlación por rangos también se utiliza para variables cuantitativas, con la aclaración de que el grado de asociación obtenido no es el de los valores de las variables, sino el de las clasificaciones por rangos de dichos valores. Este coeficiente viene dado por:

$$\rho = 1 - \frac{6\sum d_i^2}{N^3 - N}$$

siendo $d_i = x_i - y_i$. Este coeficiente también se denomina coeficiente de correlación ordinal, y por ser un coeficiente de correlación varía entre −1 y 1. Cuando la concordancia entre los rangos es perfecta, entonces $d_i = x_i - y_i = 0$ y $r = 1$. Cuando la discordancia es perfecta, $r = -1$. Cuando no hay ni concordancia ni discordancia, $r = 0$.

Matriz de correlaciones y covarianzas

Cuando se tiene una variable tridimensional (X,Y,Z) o enedimensional en general, también se puede realizar la descripción y análisis de las distribuciones de frecuencias subyacentes. La complejidad crece cuando el número de variables o factores que se analizan simultáneamente aumenta, pero conocido el procedimiento para el caso tridimensional, su generalización al n-dimensional es inmediata. El análisis es igualmente válido cuando los caracteres analizados son de naturaleza cuantitativa o cualitativa.

Al igual que en las distribuciones bidimensionales, la forma más usual de representación de distribuciones tridimensionales son las tablas de correlación (referentes a caracteres cuantitativos) o las tablas de contingencia (referentes a caracteres cuantitativos). Como sucedía en el caso de la distribución bidimensional, se pretende que las frecuencias, tanto conjuntas y marginales como condicionadas, sean fácilmente localizables, y también sus respectivas distribuciones.

Un elemento esencial en el estudio de variables enedimensionales es la *matriz de covarianzas*, que resume las covarianzas para todos los posibles pares de variables de entre *n* dadas $X_1, X_2, ..., X_n$. Se define como:

$$C=\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \cdots & \sigma_{nn} \end{pmatrix}$$

donde cada s_{ij} representa la covarianza entre X_i y X_j para todo *i,j*.

El signo de cada σ_{ij} indica el sentido de la variación conjunta de las dos variables X_i y X_j que estamos considerando. Si la covarianza es positiva, quiere decir que ambas variables varían en el mismo sentido, mientras que si la variación de las mismas tiene lugar en sentido contrario, la covarianza tomará valores negativos.

Con la matriz de covarianzas analizamos simultáneamente el sentido de la variación conjunta de todos los posibles pares de variables X_i y X_j para todo *i,j*.

Otro elemento esencial en el estudio de variables enedimensionales es la *matriz de correlaciones*, que resume las correlaciones para

FIGURA 10-8

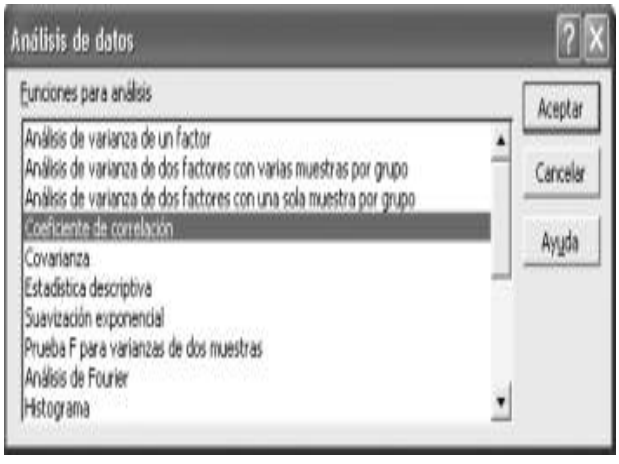


FIGURA 10-9

FIGURA 10-10

todos los posibles pares de variables de entre n dadas X_1, X_2,...,X_n. Se define como:

$$R=\begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2n} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ r_{n1} & r_{n2} & r_{n3} & \cdots & r_{nn} \end{pmatrix}$$

donde cada *rij* es el coeficiente de correlación entre X_i y X_j para todo *i,j*.

Si dada una serie de variables $X_1, X_2, \ldots, X_n$, se trata de estudiar el grado de dependencia simultánea entre todas ellas (o bien entre grupos de ellas), puede utilizarse la matriz de correlaciones. Si en base a la intensidad con que dependen se puede establecer una función que explique una variable mediante todas las demás, que se supone son sus causas influyentes, estamos ante un problema de regresión múltiple, que será estudiado en capítulos posteriores.

Mediante el *coeficiente de correlación lineal múltiple* se estudia el grado de asociación lineal simultánea entre todas las variables, mientras que mediante los coeficientes de correlación simples r_{ij} se mide el grado de asociación entre las variables X_i y X_j sin tener en cuenta a las demás variables.

	A	B	C	D
1	X	Y	Z	
2	1	5	9	
3	2	6	1	
4	3	7	2	
5	4	8	3	
6				
7		*X*	*Y*	*Z*
8	X	1		
9	Y	1	1	
10	Z	-0,61065803	-0,61065803	1

FIGURA 10-11

10.6. CORRELACIÓN MÚLTIPLE MEDIANTE HERRAMIENTAS DE ANÁLISIS

FIGURA 10-12

Excel proporciona herramientas de análisis para medir la relación entre dos conjuntos de datos. El cálculo de la correlación devuelve la covarianza de dos conjuntos de datos dividida por el producto de sus desviaciones estándar. Se puede utilizar la herramienta *Coeficiente de correlación* para determinar si dos conjuntos de datos varían conjuntamente, es decir, si los valores altos de un conjunto están asociados con los valores altos del otro (correlación positiva), si los valores bajos de un conjunto están asociados con los valores bajos

del otro (correlación negativa), o si los valores de ambos conjuntos no están relacionados (correlación con tendencia a cero). Cuando se consideran más de dos variables, esta herramienta devuelve la *matriz de correlaciones* entre ellas.

Correlación y matriz de correlaciones

FIGURA 10-13

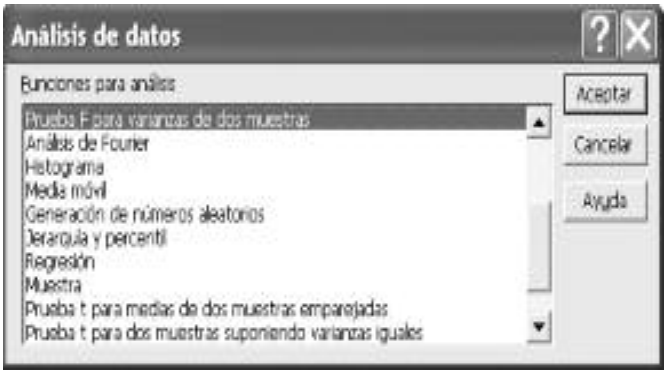


FIGURA 10-14

La opción *Análisis de datos* del menú *Herramietas* (Figura 10-8) nos lleva al cuadro de diálogo *Análisis de datos* de la Figura 10-9. Si en la lista *Funciones para análisis* elegimos *Coeficiente de correlación,* se obtiene el cuadro de diálogo de la Figura 10-10, que permite calcular la matriz de correlaciones de las variables especificadas en el campo *Rango de entrada*.

En el campo *Rango de entrada* introduzca la referencia de celda del rango de datos que desee analizar (rango que contiene las variables cuya correlación o matriz de correlaciones se va a calcular). La referencia debe-

FIGURA 10-15

rá contener dos o más rangos adyacentes organizados en columnas o filas. En el campo *Agrupado por* haga clic en el botón *Filas* o *Columnas* para indicar si los datos del rango de entrada están organizados en filas o en columnas. Si la primera fila del rango de entrada contiene rótulos, active la casilla de verificación *Rótulos en la primera fila*. Si los rótulos están en la primera columna del rango de entrada, active la casilla de verificación *Rótulos en la primera columna*. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.

En cuanto a las *Opciones de salida*, en el campo *Rango de salida* introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados (matriz de correlaciones). Microsoft Excel sólo completará media tabla, ya que la correlación entre dos rangos de datos es independiente del orden en que se procesen dichos rangos. Las celdas de la tabla de resultados con coordenadas de filas y de columnas iguales contendrán el valor 1, ya que cada conjunto de datos está perfectamente correlacionado consigo mismo. Haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado. En la Figura 10-11, se muestra la matriz de correlaciones de las variables *X, Y* y *Z* para las opciones de entrada en el cuadro *Coeficiente de correlación* de la Figura 10-12. Se observa la escasa relación existente entre las variables, ya que todos los coeficientes de correlación son muy pequeños.

Además de las funciones del menú de análisis de datos podemos también usar XLSTAT → Pruebas de correlación / asociación → Pruebas de correlación (Figura 10-13).

10.7. CONTRASTES DE HIPÓTESIS CON HERRAMIENTAS DE ANÁLISIS

Excel contiene varias herramientas de análisis útiles para realizar para realizar contrastes de hipótesis. La opción *Análisis de datos* del menú *Herramietas* (Figura 10-13) nos lleva al cuadro de diálogo de la Figura 10-14. Si en la lista *Funciones para análisis* elegimos *Prueba F para varianzas de dos muestras,* podremos realizar contrastes de igualdad de varianzas. Para realizar contrastes de igualdad de medias con varianzas iguales y desconocidas, se utiliza la opción *Prueba t para dos muestras suponiendo varianzas iguales*. Para realizar contrastes de igualdad de medias con varianzas desiguales y desconocidas, se utiliza la opción *Prueba t para dos muestras suponiendo varianzas desiguales.* Para realizar contrastes de igualdad de medias con varianzas conocidas, se utiliza la opción *Prueba z para medias de dos muestras*.

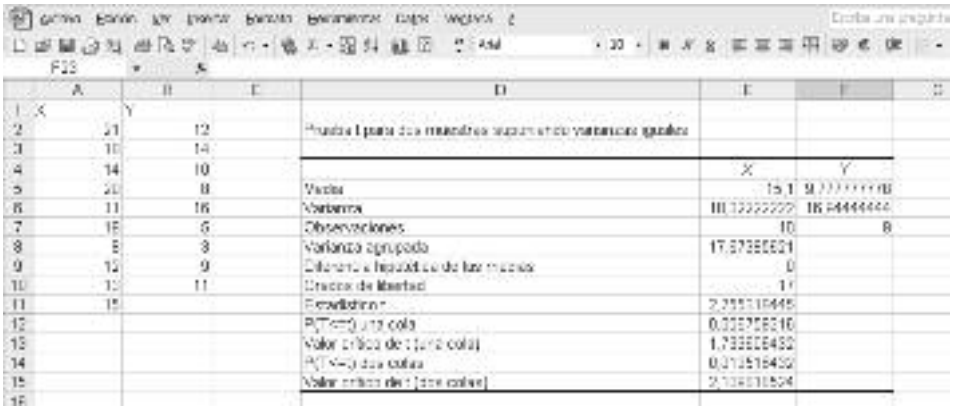


FIGURA 10-16

Contraste *T* para diferencias de medias suponiendo varianzas iguales y desconocidas

En Excel es posible ejecutar una prueba *T* de Student en dos muestras para determinar si sus medias son iguales suponiendo que las varianzas de ambos conjuntos de datos son desconocidas e iguales. Esta prueba se conoce con el nombre de *prueba t homoscedástica*. Si en el cuadro de diálogo *Análisis de datos* de la Figura 11-2 elegimos *Prueba t para dos muestras suponiendo varianzas iguales*, se obtiene el cuadro de diálogo de la Figura 10-15.

Los campos de la Figura 10-15 tienen las siguientes funcionalidades:

FIGURA 10-17

FIGURA 10-18

Rango para la variable 1: Introduzca la referencia de celda correspondiente al primer rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Rango para la variable 2: Introduzca la referencia de celda correspondiente al segundo rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Diferencia hipotética entre medias: Introduzca el número que desee para realizar el cambio en las medias de las muestras. Un valor 0 (cero) indica que, según la hipótesis, las medias de las muestras serán iguales.

Rótulos: Active esta casilla si la primera fila o la primera columna del rango de entrada contienen rótulos. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.

Alfa: Introduzca el nivel de confianza para la prueba. Este valor deberá estar comprendido en el rango 0 - 1. El nivel *Alfa* es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

Rango de salida: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Microsoft Excel determinará el tamaño del área de resultados, y mostrará un mensaje si la tabla de resultados reemplaza datos ya existentes.

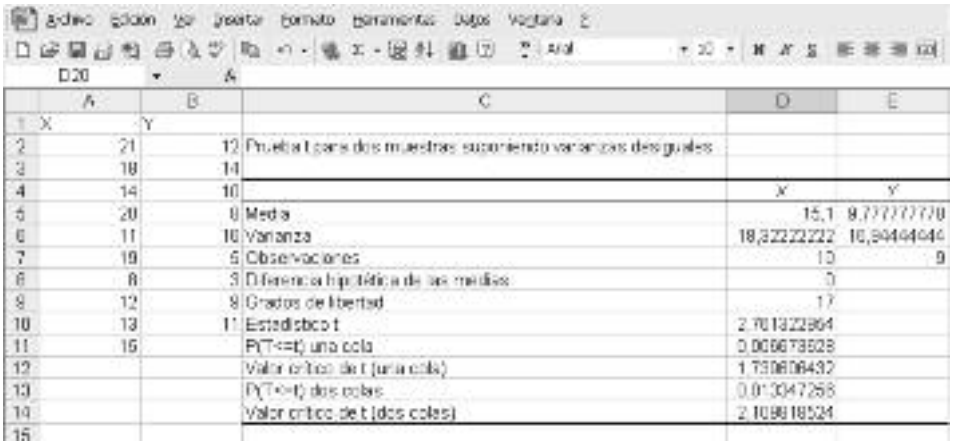
Opciones de salida: Haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

En la Figura 10-16 se muestra la salida correspondiente a las opciones de *Prueba t para dos muestras suponiendo varianzas iguales* de la Figura 10-15.

En el ejemplo de las muestras *X* e *Y* de la Figura 10-16 se rechaza la igualdad de medias, tanto para el contraste de una cola como para el contraste de dos colas, ya que ambos valores críticos de *T* (1,73960643 y 2,10981852) son menores que el valor del estadístico *t* (2,75531845), es decir, caen dentro de la región crítica o de rechazo. Además, las dos probabilidades o *p-valores* (0,00675822 y 0,01351643) son menores o iguales que *Alfa* (0,05).

Contraste *T* para diferencias de medias suponiendo varianzas desiguales y desconocidas

En Excel es posible ejecutar una prueba *T* de Student en dos muestras para determinar si sus medias son iguales, suponiendo que las varian-



	A	B	C	D	E
1	X	Y			
2	21	12	Prueba t para dos muestras suponiendo varianzas desiguales		
3	18	14			
4	14	10		X	Y
5	20	8	Media	15,1	9,777777778
6	11	10	Varianza	18,32222222	16,94444444
7	19	5	Observaciones	10	9
8	8	3	Diferencia hipotética de las medias	0	
9	12	9	Grados de libertad	17	
10	13	11	Estadístico t	2,701822864	
11	15		P(T<=t) una cola	0,006673528	
12			Valor crítico de t (una cola)	1,739606432	
13			P(T<=t) dos colas	0,013347056	
14			Valor crítico de t (dos colas)	2,109818524	
15					

FIGURA 10-19

zas de ambos conjuntos de datos son desconocidas y desiguales. Esta prueba se conoce con el nombre de *prueba t heteroscedástica*. Si en el cuadro de diálogo *Análisis de datos* de la Figura 10-17 elegimos *Prueba t para dos muestras suponiendo varianzas desiguales*, se obtiene el cuadro de diálogo de la Figura 11-6.

Los campos de la Figura 10-18 tienen las siguientes funcionalidades:

FIGURA 10-20

FIGURA 10-21

Rango para la variable 1: Introduzca la referencia de celda correspondiente al primer rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Rango para la variable 2: Introduzca la referencia de celda correspondiente al segundo rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Diferencia hipotética entre medias: Introduzca el número que desee para realizar el cambio en las medias de las muestras. Un valor 0 (cero) indica que, según la hipótesis, las medias de las muestras serán iguales.

Rótulos: Active esta casilla si la primera fila o la primera columna del rango de entrada contienen rótulos. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.

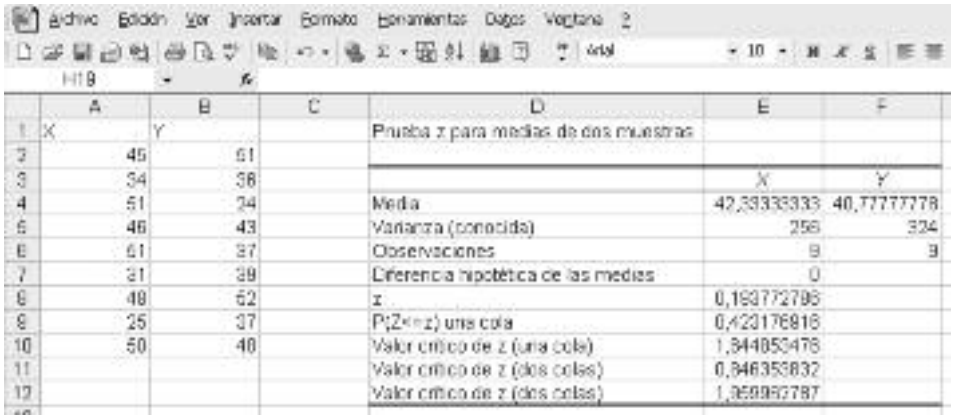
Alfa: Introduzca el nivel de confianza para la prueba. Este valor deberá estar comprendido en el rango entre 0 y 1. El nivel *Alfa* es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

Rango de salida: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Microsoft Excel determinará el tamaño del área de resultados, y mostrará un mensaje si la tabla de resultados reemplaza datos ya existentes.

Opciones de salida: Haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

En la Figura 10-19 se muestra la salida correspondiente a las opciones de *Prueba t para dos muestras suponiendo varianzas iguales* de la Figura 11-6.

En el ejemplo de las muestras *X* e *Y* de la Figura 10-19 se rechaza la igualdad de medias, tanto para el contraste de una cola como para el contraste de dos colas, ya que ambos valores críticos de *T* (1,73960643 y 2,10981852) son menores que el valor del estadístico *t* (2,76132295), es decir, caen dentro de la región crítica o de rechazo. Además las dos probabilidades o *p-valores* (0,00667363 y 0,01334726) son menores o iguales que *Alfa* (0,05).



	A	B	C	D	E	F
1	X	Y		Prueba z para medias de dos muestras		
2	45	61				
3	54	58			X	Y
4	51	54		Media	42,33333333	40,77777778
5	46	43		Varianza (conocida)	256	324
6	61	37		Observaciones	9	9
7	31	39		Diferencia hipotética de las medias	0	
8	48	62		z	0,193772795	
9	25	37		P(Z<=z) una cola	0,423178816	
10	50	48		Valor crítico de z (una cola)	1,644853476	
11				Valor crítico de z (dos colas)	0,846356832	
12				Valor crítico de z (dos colas)	1,959963787	

FIGURA 10-22

Contraste *Z* para diferencias de medias suponiendo varianzas conocidas

En Excel es posible ejecutar una prueba *Z* de la normal en dos muestras para determinar si sus medias son iguales, suponiendo que las varianzas de ambos conjuntos de datos son conocidas. Si en el cuadro de diálogo *Análisis de datos* de la Figura 10-20 elegimos *Prueba Z para medias de dos muestras,* se obtiene el cuadro de diálogo de la Figura 10-21.

Los campos de la Figura 10-21 tienen las siguientes funcionalidades:

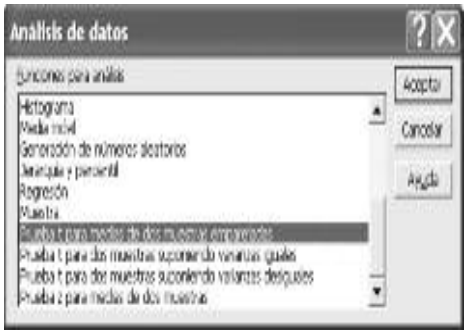


FIGURA 10-23

FIGURA 10-24

Rango para la variable 1: Introduzca la referencia de celda correspondiente al primer rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Rango para la variable 2: Introduzca la referencia de celda correspondiente al segundo rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Diferencia hipotética entre medias: Introduzca el número que desee para realizar el cambio en las medias de las muestras. Un valor 0 (cero) indica que, según la hipótesis, las medias de las muestras serán iguales.

Varianza para las variables 1 y 2: Se introducen dichas varianzas conocidas.

Rótulos: Active esta casilla si la primera fila o la primera columna del rango de entrada contienen rótulos. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.

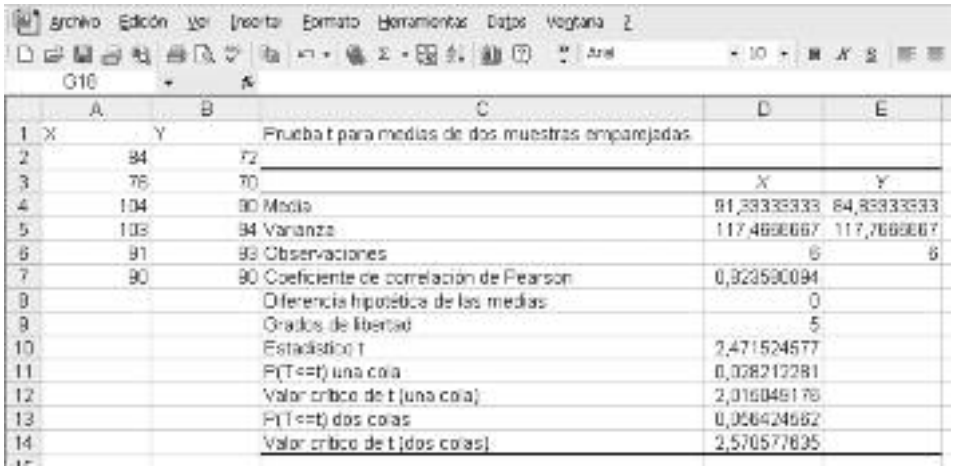
Alfa: Introduzca el nivel de confianza para la prueba. Este valor deberá estar comprendido en el rango 0 - 1. El nivel *Alfa* es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

Rango de salida: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Microsoft Excel determinará el tamaño del área de resultados, y mostrará un mensaje si la tabla de resultados reemplaza datos ya existentes.

Opciones de salida: Haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

En la Figura 10-22 se muestra la salida correspondiente a las opciones de *Prueba Z para medias de dos muestras* de la Figura 10-21.

En el ejemplo de las muestras *X* e *Y* de la Figura 10-22, se acepta la igualdad de medias, tanto para el contraste de una cola como para el contraste de dos colas, ya que ambos valores críticos de *Z* (1,644853 y 1,95996108) son mayores que el valor del estadístico *Z* (0,19377279), es



	A	B	C	D	E
1	X	Y	Prueba t para medias de dos muestras emparejadas		
2	84	72			
3	76	70		X	Y
4	104	90	Media	91,33333333	84,83333333
5	103	84	Varianza	117,4666667	117,7666667
6	91	93	Observaciones	6	6
7	90	90	Coeficiente de correlación de Pearson	0,922590094	
8			Diferencia hipotética de las medias	0	
9			Grados de libertad	5	
10			Estadístico t	2,471524577	
11			P(T<=t) una cola	0,028212281	
12			Valor crítico de t (una cola)	2,015048176	
13			P(T<=t) dos colas	0,056424562	
14			Valor crítico de t (dos colas)	2,570577835	

FIGURA 10-24

decir, caen fuera de la región crítica o de rechazo. Además, las dos probabilidades o *p-valores* son mayores o iguales que *Alfa* (0,05).

Contraste *T* para diferencias de medias en muestras pareadas con varianzas desiguales y desconocidas

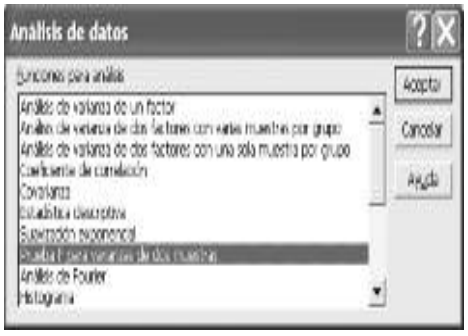


FIGURA 10-25

FIGURA 10-26

En Excel es posible ejecutar una prueba *T* de Student en dos muestras pareadas para determinar si las medias de las dos muestras son iguales suponiendo que las varianzas de ambos conjuntos de datos son desiguales. Puede utilizarse una prueba pareada cuando haya un par natural de observaciones en las muestras (como cuando un grupo de muestra se so-

mete dos veces a prueba, antes de un experimento y después de éste). Si en el cuadro de diálogo *Análisis de datos* de la Figura 10-23 elegimos *Prueba T para medias de dos muestras emparejadas*, se obtiene el cuadro de diálogo de la Figura 10-24.

Los campos de la Figura 10-24 tienen las siguientes funcionalidades:

Rango para la variable 1: Introduzca la referencia de celda correspondiente al primer rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Rango para la variable 2: Introduzca la referencia de celda correspondiente al segundo rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Diferencia hipotética entre medias: Introduzca el número que desee para realizar el cambio en las medias de las muestras. Un valor 0 (cero) indica que, según la hipótesis, las medias de las muestras serán iguales.

Rótulos: Active esta casilla si la primera fila o la primera columna del rango de entrada contienen rótulos. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.

Alfa: Introduzca el nivel de confianza para la prueba. Este valor deberá estar comprendido en el rango entre 0 - 1. El nivel *Alfa* es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

Rango de salida: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Microsoft Excel determinará

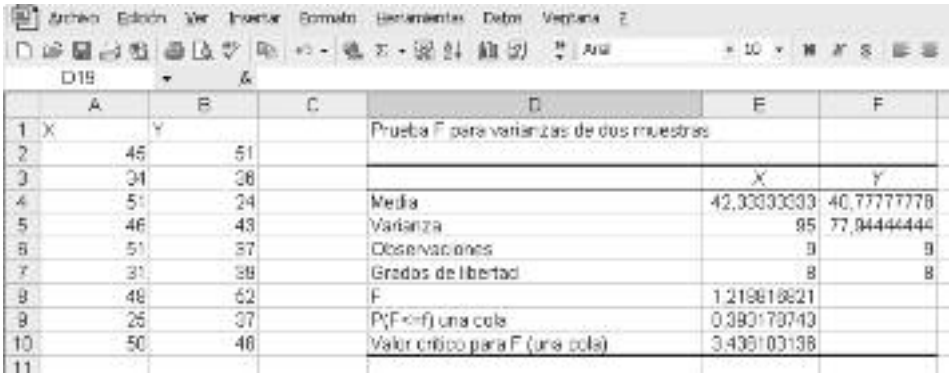


FIGURA 10-27

el tamaño del área de resultados, y mostrará un mensaje si la tabla de resultados reemplaza datos ya existentes.

En cuanto a las *Opciones de salida*, haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado. En la Figura 10-25 se muestra la salida correspondiente a las opciones de *Prueba t para medias de dos muestras emparejadas* de la Figura 10-24.

EJERCICIO 10-1. Sean las variables X, Y y Z, cuyos valores son los siguientes:

X	Y	Z
2	4	2
3	5	4
6	10	6
8	11	7
10	15	10

a) **Calcular la media, la mediana, la desviación típica, la varianza y los coeficientes de asimetría y curtosis para las tres variables, y hallar un intervalo de confianza para la media basado en cada variable con un nivel de confianza del 5% (coeficiente de confianza del 95%).**

b) **Hallar la matriz de correlaciones y deducir el grado de dependencia entre las variables.**

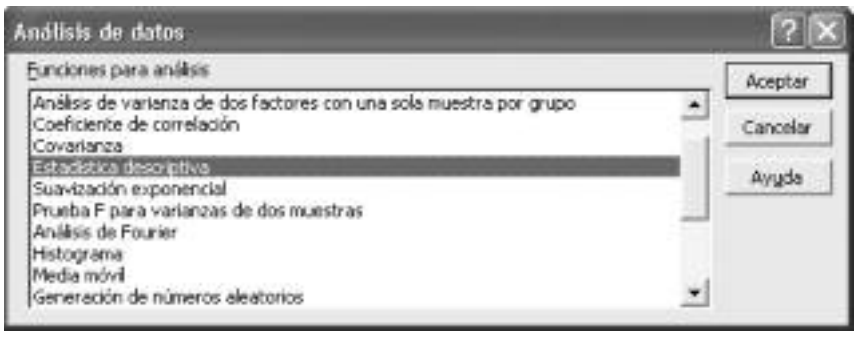


FIGURA 10-28

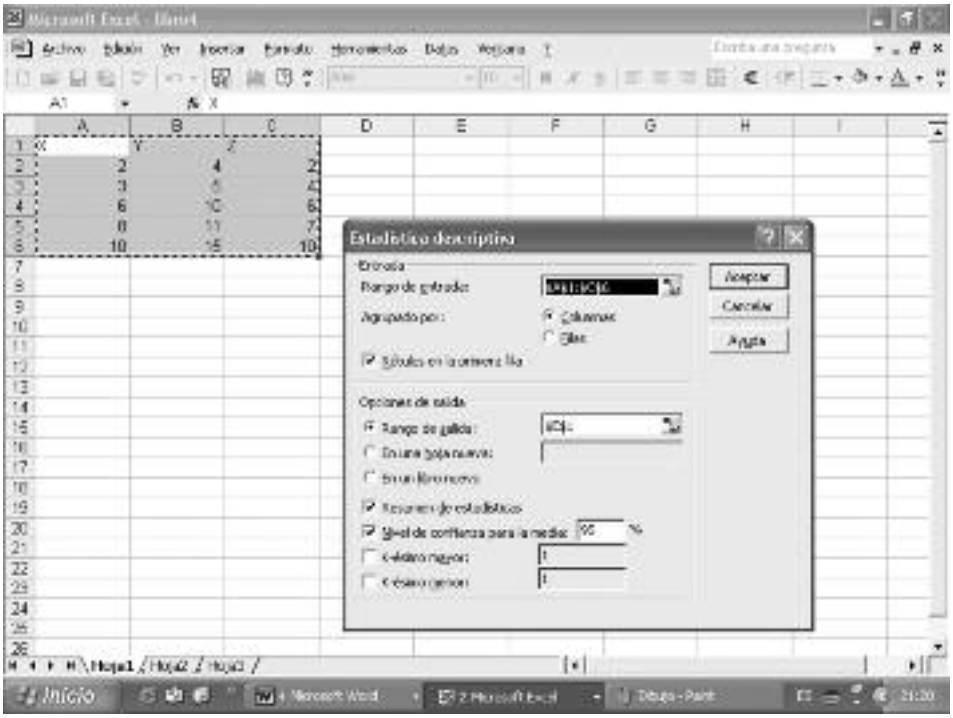


FIGURA 10-29

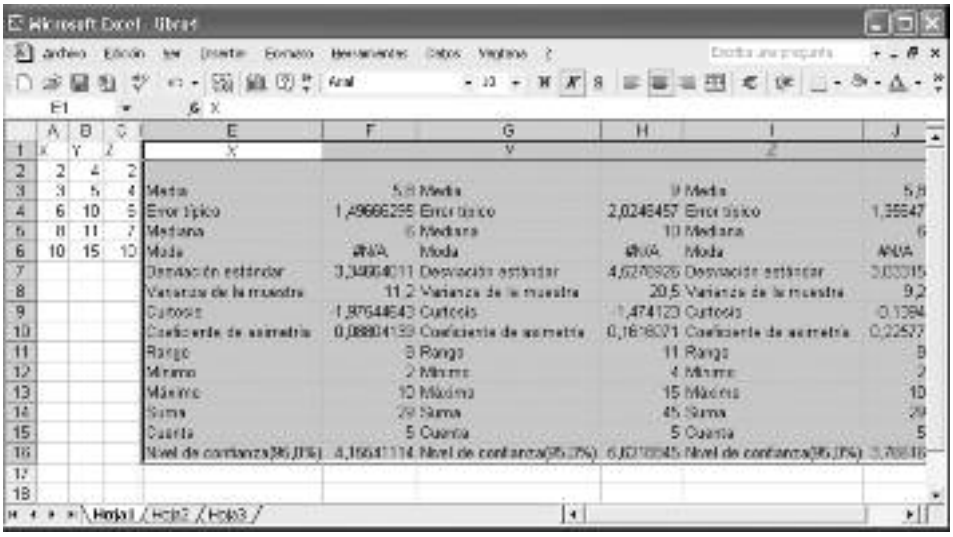


FIGURA 10-30

En el ejemplo de las muestras emparejadas *X* e *Y* de la Figura 11-13 se rechaza la igualdad de medias para el contraste de una cola, ya que el valor crítico de *T* (2,01504918) es menor que el valor del estadístico *T* (2,47152458), es decir, cae en la región crítica o de rechazo. Además, la pro-

FIGURA 10-31

FIGURA 10-32

L	M	N	O
	X	*Y*	*Z*
X	1		
Y	0,9899319	1	
Z	0,98021232	0,98302129	1

FIGURA 10-33

babilidad o *p-valor* (0,02821228) es menor o igual que *Alfa* (0,05). Sin embargo, se acepta la igualdad de medias para el contraste de dos colas, ya que el valor crítico de *t* (2,57057764) es mayor que el valor del estadístico *t*

EJERCICIO 10-2. El consumo (C) y la renta mensual (RM) de 100 familias, expresadas en 104 euros, se presentan en la siguiente tabla bidimensional de frecuencias:

C / RM	15	25	35	45
30	10	15		
40	5	20	25	
50		15	5	5

a) **Hallar la relación lineal subyacente entre el consumo y la renta.**
b) **Cuantificar el grado de representatividad de la relación lineal anterior.**
c) **Hallar el consumo esperado para una renta de $60*10^4$ euros.**
d) **Hallar las distribuciones marginales de las variables C y RM y su medias, varianzas, desviaciones típicas y coeficientes de asimetría y curtosis.**

(2,47152458), es decir, cae fuera de la región crítica o de rechazo. Además, la probabilidad o *p-valor* (0,05642456) es mayor o igual que *Alfa* (0,05).

Contraste *F* para igualdad de varianzas

En Excel es posible ejecutar una prueba *F* de Fisher Snedocor para determinar si las varianzas de las dos muestras son iguales (cociente de varianzas igual a la unidad). Si en el cuadro de diálogo *Análisis de datos* de la Figura 10-25 elegimos *Prueba F para varianzas de dos muestras,* se obtiene el cuadro de diálogo de la Figura 10-26.

Los campos de la Figura 10-26 tienen las siguientes funcionalidades:

Rango para la variable 1: Introduzca la referencia de celda correspondiente al primer rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

Rango para la variable 2: Introduzca la referencia de celda correspondiente al segundo rango de datos que desee analizar. El rango debe constar de una única columna o una única fila de datos.

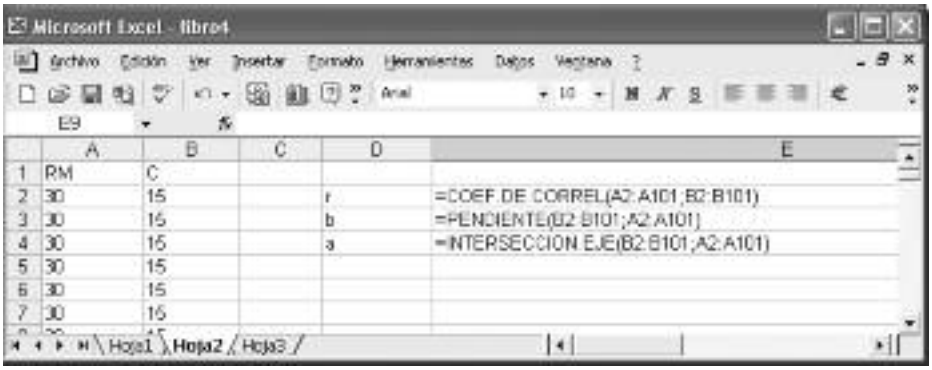


FIGURA 10-34

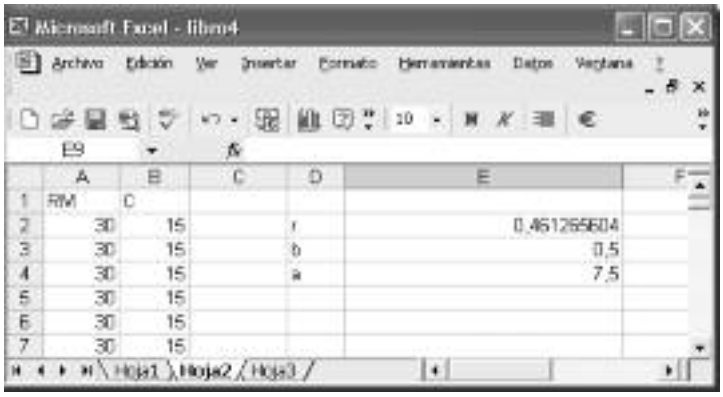


FIGURA 10-35

Rótulos: Active esta casilla si la primera fila o la primera columna del rango de entrada contienen rótulos. Esta casilla de verificación estará desactivada si el rango de entrada carece de rótulos; Microsoft Excel generará los rótulos de datos correspondientes para la tabla de resultados.

Alfa: Introduzca el nivel de confianza para la prueba. Este valor deberá estar comprendido en el rango 0 - 1. El nivel *Alfa* es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

En el cuadro *Rango de salida* introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Microsoft Excel determinará el tamaño del área de resultados, y mostrará un mensaje si la tabla de resultados reemplaza datos ya existentes.

En cuanto a las *Opciones de salida*, haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para darle

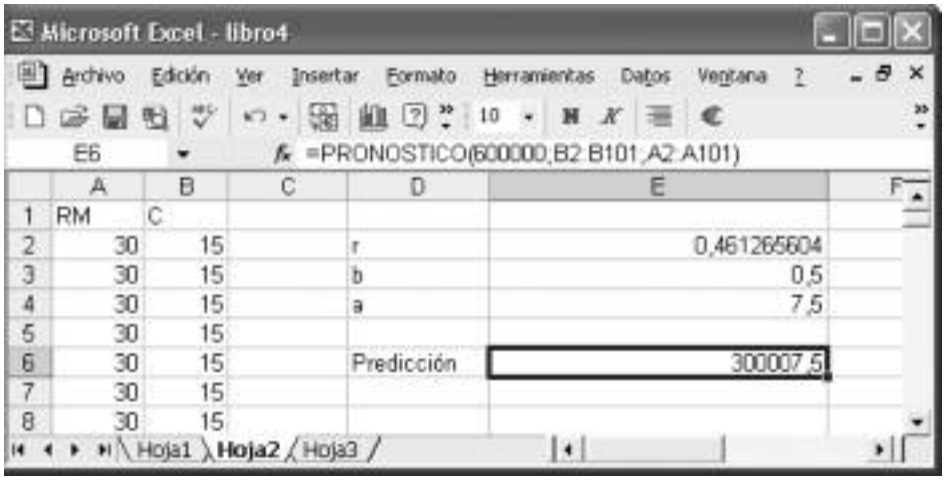


FIGURA 10-36

un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

En la Figura 10-27 se muestra la salida correspondiente a las opciones de *Prueba F para varianzas de dos muestras* de la Figura 10-26.

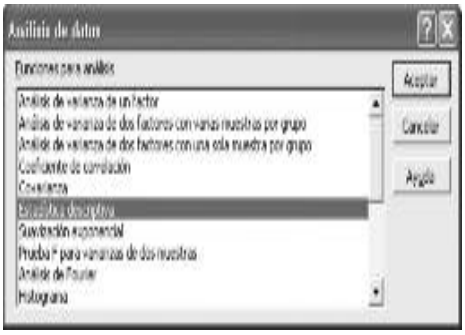


FIGURA 10-37

FIGURA 10-38

	G	H	I	J
	AM		C	
	Media	40	Media	27,5
	Error típico	0,71086905	Error típico	0,7703469
	Mediana	40	Mediana	26
	Moda	40	Moda	25
	Desviación estándar	7,10869065	Desviación estándar	7,7034688
	Varianza de la muestra	50,5050505	Varianza de la muestra	59,3434343
	Curtosis	-0,9693751	Curtosis	-0,2527384
	Coeficiente de asimetría	4,5773E-17	Coeficiente de asimetría	0,2113737
	Rango	30	Rango	30
	Mínimo	30	Mínimo	16
	Máximo	60	Máximo	46
	Suma	4000	Suma	2750
	Cuenta	100	Cuenta	100

FIGURA 10-39

En el ejemplo de las muestras *X* e *Y* de la Figura 11-16 se acepta la igualdad de varianzas para el contraste de una cola, ya que el valor crítico de *F* (3,438110314) es mayor que el valor del estadístico *F* (1,21881682), es decir, cae fuera de la región crítica o de rechazo. Además la probabilidad o *p-valor* (0,39317874) es mayor o igual que *Alfa* (0,05).

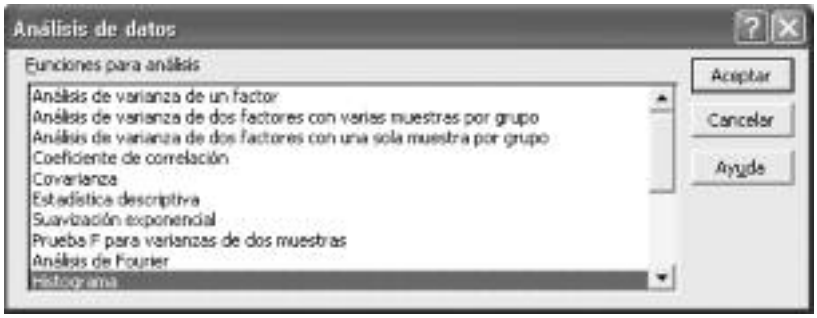


FIGURA 10-40

FIGURA 10-41

FIGURA 10-42

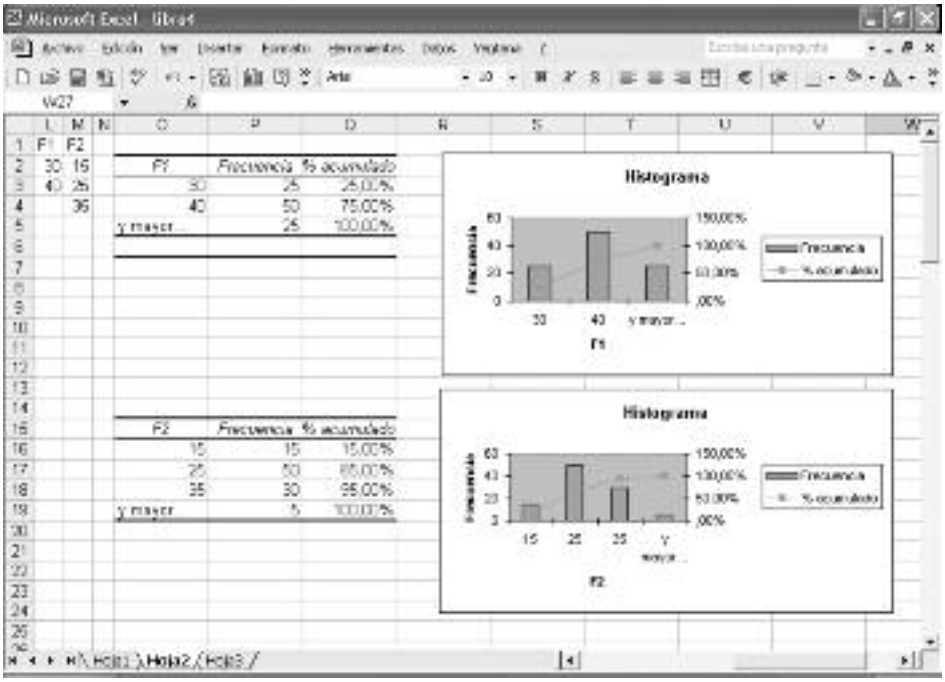


FIGURA 10-43

EJERCICIO 10-3. Se sospecha que existe una diferencia significativa entre la proporción de hombres y la proporción de mujeres que acceden a las ayudas a la familia de una población. Para salir de dudas al 95% de certidumbre, se toma una muestra aleatoria de 300 hombres, de los cuales 27 reciben o han recibido la ayuda. Asimismo tomamos una muestra de 400 mujeres, de las cuales 32 reciben o han recibido la ayuda. ¿Qué nos indican estos resultados?

Comenzamos introduciendo en la hoja de cálculo las variables *X, Y* y *Z* como columnas de la misma. A continuación, seleccionamos la opción *Análisis de datos* del menú *Herramientas,* y elegimos *Estadística descriptiva* en *Funciones para análisis* (Figura 10-28). Rellenamos la pantalla *Estadística descriptiva* como se indica en la Figura 10-29. Al pulsar *Aceptar,* se obtienen los resultados de la Figura 10-30.

Para calcular la matriz de correlaciones seleccionamos la opción *Análisis de datos* del menú *Herramientas,* y elegimos *Coeficiente de correlación* en *Funciones para análisis* (Figura 10-31). Rellenamos la pantalla *Coeficiente de correlación* como se indica en la Figura 10-32. Al pulsar *Aceptar,* se obtiene la matriz de correlaciones de la Figura 10-33.

En la salida anterior, se observa que el coeficiente de correlación entre las variables *X* e *Y* es 0,9899, entre *X* y *Z* 0,9802, y entre *Y* y *Z* 0,9830, lo que indica la fuerte dependencia lineal positiva entre cada par de variables.

En este caso, observamos que los valores de las variables vienen dados en forma de tabla bidimensional de frecuencias absolutas. La primera tarea será introducir adecuadamente los valores de las dos variables como dos columnas de la hoja de cálculo. El par de valores *RM=30* y *C=15* se introduce 10 veces, el par de valores *RM=40 C=15* se introduce 5 veces, y así sucesivamente; se introduce cada par de valores tantas veces como indique su frecuencia absoluta.

FIGURA 10-44

	A	B	C
2	0,46778027	2,24140146	0,63994175
3			

FIGURA 10-45

Es lógico que el consumo de las familias dependa de su renta mensual, con lo que hemos de hallar la función *f* tal que *C=f(RM),* es decir, la recta de regresión de *C* sobre *RM*. Una vez calculada la ecuación de esta recta, la utilizaremos para predecir el consumo esperado para una renta de $60*10^4$ euros.

Una vez introducidos los datos de las variables, utilizamos las funciones de Excel que permiten calcular el coeficiente de correlación y la pen-

EJERCICIO 10-4. Las ayudas en euros recibidas por las familias de dos ayuntamientos de características similares son:

Grupo 1.º: 100; 102; 96; 106; 110; 110; 120; 112; 112; 90

Grupo 2.º: 104; 88; 100; 98; 102; 92; 96; 100; 96; 96

a) **Suponiendo que las dos poblaciones son normales y de varianzas iguales y desconocidas, contrastar la hipótesis de igualdad de medias al nivel 0,05.**

b) **Realizar el contraste para varianzas desiguales.**

c) **Realizar el contraste para muestras pareadas.**

d) **Contrastar también la igualdad de varianzas.**

diente y ordenada en el origen de la recta de regresión de *C* sobre *RM* (Figura 10-34). La Figura 10-35 presenta los resultados.

Se observa que el coeficiente de correlación entre *C* y *RM* es 0,4613, que no es un valor lo suficientemente alto como para asegurar una dependencia fuerte entre las dos variables.

No obstante, correlación débil no implica independencia, por lo que puede hallarse la recta de regresión de *C* sobre *RM*, pero con la precaución de que dicha relación lineal entre ambas variables puede no ser buena, y sobre todo pueden no ser fiables las predicciones basadas en dicha relación de linealidad. Una vez calculadas la pendiente y la ordenada en el origen de la recta de regresión, tenemos la relación *C=7,5+0,5RM*.

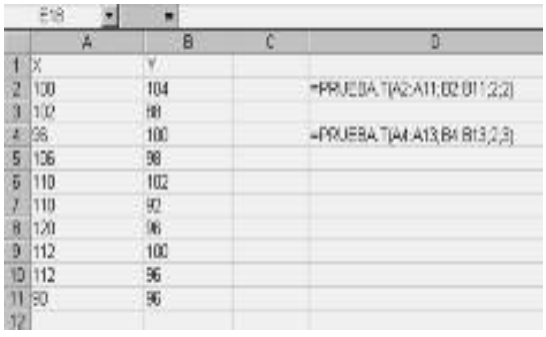


FIGURA 10-46

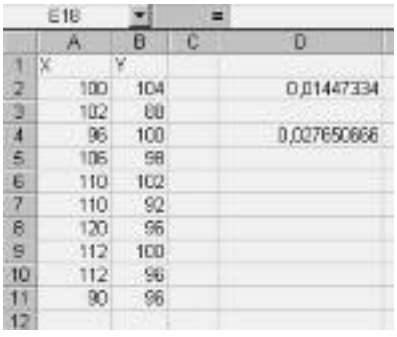


FIGURA 10-47

Una vez calculada la ecuación de la recta de regresión, la utilizaremos para predecir el consumo esperado relativo a una producción de $60*10^4$ euros, que vendrá dado por $C=7,5+0,5*60*10^4 = 37,5*10^4$ euros. Esta predicción puede no ser lo suficientemente buena, ya que el contraste de ajuste por regresión aceptaba la incorrelación entre las variables, y el coeficiente de correlación no es muy próximo a la unidad (0,4949).

FIGURA 10-48

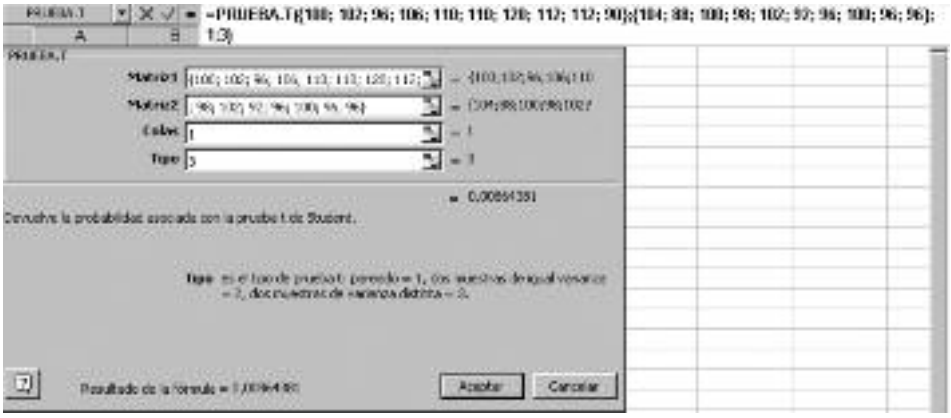


FIGURA 10-49

La predicción también puede calcularse mediante la función PRONOSTICO de Excel, tal y como se indica en la Figura 10-36.

Para hallar las medias, varianzas, desviaciones típicas y coeficientes de variación, asimetría y curtosis de las distribuciones marginales de *C* y *RM*, seleccionamos la opción *Análisis de datos* del menú *Herramientas,* y elegimos *Estadística descriptiva* en *Funciones para análisis* (Figura 10-37). Rellenamos la pantalla *Estadística descriptiva* como se indica en la Figura 10-38. Al pulsar *Aceptar,* se obtienen los resultados de la Figura 10-39.

Para hallar las propias distribuciones marginales de *C* y *RM*, seleccionamos la opción *Análisis de datos* del menú *Herramientas,* y elegimos *Histograma* en *Funciones para análisis* (Figura 10-40). Rellenamos la pantalla *Histograma* como se indica en la Figura 10-41. Al pulsar *Aceptar,* se obtiene la distribución e histograma de *RM*. Repitiendo el proceso para *C* (Figura 10-42), se obtiene la marginal de *C*. Los resultados se observan en la Figura 10-43.

Realizaremos un contraste de igualdad de proporciones (diferencia de proporciones nula) de dos poblaciones binomiales, siendo las pro-

FIGURA 10-50

FIGURA 10-51

	A	B	C	D	E	F
1	G1	G2	Prueba t para dos muestras suponiendo varianzas iguales			
2	100	104				
3	102	88		G1	G2	
4	96	100	Media	105,8	97,2	
5	106	98	Varianza	78,6222222	22,4	
6	110	102	Observaciones	10	10	
7	110	92	Varianza agrupada	50,5111111		
8	120	96	Diferencia hipotética de las medias	0		
9	112	100	Grados de libertad	18		
10	112	96	Estadístico t	2,70576449		
11	90	96	P(T<=t) una cola	0,00723667		
12			Valor crítico de t (una cola)	1,73406306		
13			P(T<=t) dos colas	0,01447334		
14			Valor crítico de t (dos colas)	2,10092367		

FIGURA 10-52

	F	G	H
1	Prueba t para dos muestras suponiendo varianzas desiguales		
2			
3		G1	G2
4	Media	105,8	97,2
5	Varianza	78,6222222	22,4
6	Observaciones	10	10
7	Diferencia hipotética de las medias	0	
8	Grados de libertad	14	
9	Estadístico t	2,70576449	
10	P(T<=t) una cola	0,00953146	
11	Valor crítico de t (una cola)	1,76130925	
12	P(T<=t) dos colas	0,01706291	
13	Valor crítico de t (dos colas)	2,1447866	

FIGURA 10-53

FIGURA 10-54

	I	J	K
1	Prueba t para medias de dos muestras emparejadas		
2			
3		G1	G2
4	Media	105,8	97,2
5	Varianza	78,6222222	22,4
6	Observaciones	10	10
7	Coeficiente de correlación de Pearson	-0,05718927	
8	Diferencia hipotética de las medias	0	
9	Grados de libertad	9	
10	Estadístico t	2,64368686	
11	P(T<=t) una cola	0,0133758	
12	Valor crítico de t (una cola)	1,83311386	
13	P(T<=t) dos colas	0,02675161	
14	Valor crítico de t (dos colas)	2,26215889	

FIGURA 10-55

porciones muestrales 27/300 = 0,09 y 32/400 = 0,08. Se utilizará el estadístico:

$$Zo = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

La región crítica del contraste será $|Z_{a/2}| > k$. Por lo tanto, para realizar el contraste al 95% bastará con calcular el valor de *Zo*, y el valor de *k* tal que *P(|Z| > k) = a/2,* con *a* = 0,05 y *Z* normal (0,1).

Tendremos presente que *P(|Z| > k) = a/2* es equivalente a *P(Z < k) = 1 - a/4*. Una vez calculados *k* y *Zo*, se comprobará si *Zo* cae dentro o fuera de la región crítica *|Z| > k*. Si cae dentro se rechaza la hipótesis nula de $p_1 - p_2 = 0$, y si cae fuera se acepta dicha hipótesis nula. Otra alternativa es calcular el p-valor del contraste *P(|Z| < Zo)* y comprobar si es mayor que

FIGURA 10-56

	L	M	N
1	Prueba F para varianzas de dos muestras		
2			
3		G1	G2
4	Media	105,8	97,2
5	Varianza	78,62222222	22,4
6	Observaciones	10	10
7	Grados de libertad	9	9
8	F	3,50992063	
9	P(F<=f) una cola	0,03769031	
10	Valor crítico para F (una cola)	3,17889715	

FIGURA 10-57

EJERCICIO 10-5. Un centro de ayuda a la familia recibe diariamente individuos nativos e inmigrantes de un nivel económico similar que denominaremos, X (nativos) e Y(inmigrantes). Con el fin de estudiar la calidad de la alimentación de sus familias, se extraen dos muestras, una de cada tipo de población, y se analiza el contenido de materia grasa ingerido, obteniendo los siguientes resultados:

X → 0,32; 0,29; 0,30; 0,28; 0,33; 0,31; 0,30; 0,29; 0,33; 0,32; 0,30; 0,29.

Y → 0,28; 0,30; 0,32; 0,29; 0,31; 0,29; 0,33; 0,32; 0,29; 0,32; 0,31; 0,29; 0,32; 0,31; 0,32; 0,33.

Realizar el contraste de hipótesis de igualdad de varianzas.

0,05, en cuyo caso se acepta la hipótesis nula de $p_1 - p_2 = 0$. Los cálculos de *Zo, k* y el p-valor se realizan en Excel mediante las fórmulas de la Figura 10-44. Los resultados se ofrecen en la Figura 10-45.

Se observa que *Zo* = 0,467 resulta menor que *k* = 2,2414, y que por lo tanto cae fuera de la región crítica. Se acepta la hipótesis nula $p_1 - p_2 = 0$ (proporciones iguales).

Además, el p-valor del contraste es 0,63, que es mayor que 0,05, lo que nos lleva a aceptar la hipótesis nula de igualdad de proporciones.

Comenzaremos introduciendo los datos de las dos muestras en dos variables (columnas) llamadas *X* e *Y* de una hoja Excel. Para contrastar la hipótesis de la igualdad de medias con varianzas iguales (2 colas) utiliza-

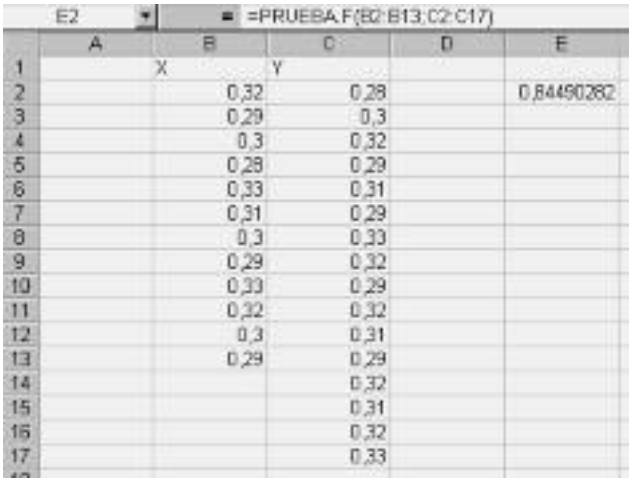


FIGURA 10-58

FIGURA 10-59

mos la función PRUEBA.T de Excel, con la sintaxis que se expone en la celda D2 de la hoja de la Figura 10-46.

Se observa que la probabilidad de que las medias coincidan es 0,0144 (Figura 10-47), lo que nos lleva a rechazar la hipótesis de igualdad de medias.

Para contrastar la hipótesis de la igualdad de medias con varianzas desiguales utilizamos la función PRUEBA.T de Excel, con la sintaxis que se expone en la celda D4 de la hoja de la Figura 11-29. Se observa que la probabilidad de que las medias coincidan es de 0,027 (Figura 10-47), lo que nos lleva a rechazar la hipótesis de igualdad de medias.

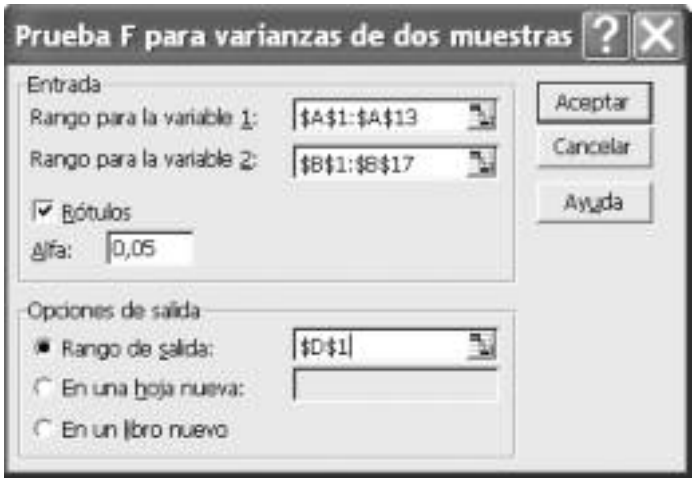


FIGURA 10-60

	A	B	C	D	E	F
1	X	Y		Prueba F para varianzas de dos muestras		
2	0,32	0,28				
3	0,29	0,3			X	Y
4	0,3	0,32		Media	0,305	0,308125
5	0,26	0,29		Varianza	0,000201818	0,00025625
6	0,33	0,31		Observaciones	12	16
7	0,31	0,29		Grados de libertad	11	15
8	0,3	0,33		F	1,089770271	
9	0,29	0,32		P(F<=f) una cola	0,42345141	
10	0,33	0,29		Valor crítico para F (una cola)	2,603896883	
11	0,32	0,32				
12	0,3	0,31				
13	0,29	0,29				
14		0,32				
15		0,31				
16		0,32				
17		0,33				
18						

FIGURA 10-61

EJERCICIO 10-6. La asociación de padres de un colegio público desea comparar las notas de sus hijos con las que obtuvieron los alumnos que estudiaron en la misma escuela el año pasado. Para ello, se toman 10 alumnos al azar de cada curso con idénticas características. Las calificaciones obtenidas son:

Escuela A → 57 49 60 55 57 48 50 61 52 56

Escuela B → 55 48 58 56 54 48 52 56 50 58

¿Podemos aceptar la homogeneidad de producciones para ambas clases de trigo a un nivel a = 0,05 suponiendo distribución normal bivariante?

También se puede realizar el contraste (esta vez para una cola) haciendo clic en el botón *Pegar función*, seleccionando el grupo de funciones *Estadísticas* y eligiendo la función PRUEBA.T, cuyas paletas se presenta adecuadamente rellenas en las Figuras 10-48 y 10-49, en cuya parte inferior se ven las probabilidades de que las varianzas no sean significativamente diferentes.

Este problema puede resolverse de forma más precisa utilizando las herramientas de análisis de Excel. Comenzamos introduciendo en una hoja de cálculo de Excel las variables G1 y G2 como columnas de la misma. A continuación, en el cuadro de diálogo *Análisis de datos* (menú *Herramientas*) elegimos *Prueba t para dos muestras suponiendo varianzas iguales,* y re-

FIGURA 10-62

FIGURA 10-63

	A	B	C	D	E	F
1	X	Y		Prueba t para medias de dos muestras emparejadas		
2	57	55				
3	43	48			X	Y
4	61	50		Media	54,5	53,5
5	55	58		Varianza	32,72222222	14,5
6	67	54		Observaciones	10	10
7	43	48		Coeficiente de correlación de Pearson	0,889540693	
8	61	52		Diferencia hipotética de las medias	0	
9	61	58		Grados de libertad	9	
10	52	50		Estadístico t	1,388757212	
11	58	58		P(T<=t) una cola	0,097584503	
12				Valor crítico de t (una cola)	1,833113856	
13				P(T<=t) dos colas	0,195169019	
14				Valor crítico de t (dos colas)	2,262158887	
15						

FIGURA 10-64

llenamos su cuadro de diálogo como se indica en la Figura 10-50. Al pulsar *Aceptar*, se obtiene la Figura 10-51, con los resultados.

Si en el cuadro de diálogo *Análisis de datos* elegimos *Prueba t para dos muestras suponiendo varianzas desiguales* y rellenamos su cuadro de diálogo como se indica en la Figura 10-52, al pulsar *Aceptar* se obtiene la Figura 10-53, con los resultados.

Si en el cuadro de diálogo *Análisis de datos* elegimos *Prueba t para medias de dos muestras emparejadas* y rellenamos su cuadro de diálogo como se indica en la Figura 10-54, al pulsar *Aceptar* se obtiene la Figura 10-55, con los resultados.

Se observa que en todos los tipos de contraste los *p-valores* son menores que 0,05, lo que implica que la diferencia de medias es estadísticamente significativa, es decir, el rechazo de la hipótesis nula de igualdad de

EJERCICIO 10-7. En las puntuaciones recibidas por familias demandantes de becas de comedor éstas recibieron las siguientes calificaciones:

80, 70, 90, 75, 55, 80, 80, 65, 100, 75, 60, 60,75, 95, 80, 80, 90, 85, 70, 95, 75, 70, 85, 80, 80, 65, 65, 50, 75, 75, 85, 85, 90, 70.

Compruébese si las puntuaciones fueron o no distribuidas según una ley normal a un nivel 0,05.

FIGURA 10-65

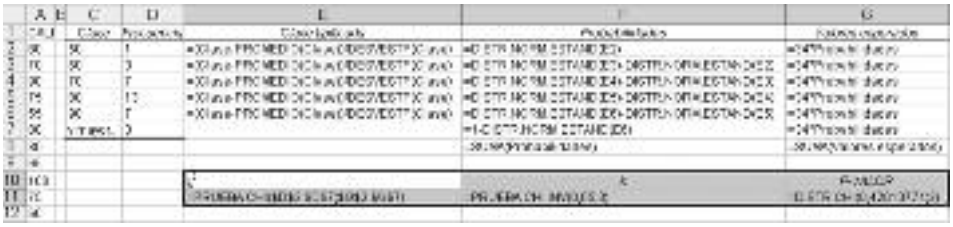


FIGURA 10-66

medias. Por otra parte, todos los valores críticos, tanto para una cola como para dos colas, son menores que el valor del estadístico, lo que corrobora la aceptación de la significatividad de la diferencia de medias para todos los tipos de contrastes.

	A	B	C	D	E	F	G
1	CALI		*Clase*	*Frecuencia*	*Clase tipificada*	*Probabilidades*	*Valores esperados*
2	80		60	1	-1,414213562	0,079849863	2,87400819
3	70		60	3	-0,707106781	0,161100334	5,47741137
4	90		70	7	0	0,260250013	8,848500433
5	75		80	13	0,707106781	0,260250013	8,848500448
6	55		90	7	1,414213562	0,161100334	5,47741137
7	90		y mayor...	3		0,079849863	2,87400819
8	80					1	34
9	65						
10	100				χ^2	*k*	*P-VALOR*
11	75				0,420107745	7,814724703	0,93805809
12	60						

FIGURA 10-67

Para contrastar la igualdad de varianzas, en el cuadro de diálogo *Análisis de datos* elegimos *Prueba F para varianzas de dos muestras,* y rellenamos su cuadro de diálogo como se indica en la Figura 10-56. Al pulsar *Aceptar*, se obtiene la Figura 10-57, con los resultados.

EJERCICIO 10-8. Repetir los ejercicios anteriores con XLSTAT

Se observa que el *p*-valor es menor que 0,05, lo que implica que el cociente de varianzas es significativamente distinto de la unidad, o sea, el rechazo de la igualdad de varianzas. Por otra parte, el valor crítico para una cola es menor que el valor del estadístico, lo que corrobora la aceptación de la hipótesis de varianzas distintas al 95% de confianza.

Comenzaremos introduciendo los datos de las dos muestras como dos variables (columnas) llamadas *X* e *Y* de una hoja Excel. Para contrastar la hipótesis de la igualdad de varianzas, utilizamos la función PRUEBA.F de Excel, con la sintaxis que se expone en la Figura 10-58. Se observa que la probabilidad de que las varianzas coincidan es 0,84, lo que nos lleva a aceptar la hipótesis de igualdad de varianzas.

También se puede hallar el resultado haciendo clic en el botón pegar función, seleccionando el grupo de funciones *Estadísticas* y eligiendo la función PRUEBA.F, cuya paleta se presenta adecuadamente rellena en la Figura 10-59, en cuya parte inferior se ve la probabilidad de que las varianzas no sean significativamente diferentes. Este resultado se inserta en la celda activa al pulsar el botón *Aceptar*.

Para contrastar la igualdad de varianzas formalmente utilizando las herramientas de análisis, en el cuadro de diálogo *Análisis de datos* elegimos *Prueba F para varianzas de dos muestras,* y rellenamos su cuadro de diálogo como se indica en la Figura 10-60. Al pulsar *Aceptar*, se obtiene la Figura 10-61, con los resultados.

Se observa que el *p-valor* es mayor que 0,05, lo que implica que el cociente de varianzas es significativamente igual a la unidad, o sea, la aceptación de la igualdad de varianzas. Por otra parte, el valor crítico para una cola es mayor que el valor del estadístico, lo que corrobora la aceptación de la hipótesis de varianzas iguales al 95% de confianza.

Comenzaremos introduciendo los datos de las dos muestras como dos variables (columnas) llamadas *X* e *Y* de una hoja del programa. Realizaremos el contraste de hipótesis de igualdad de calificaciones medias teniendo en cuenta que las dos muestras son dependientes, porque ambas están tomadas en la misma escuela y dependen de sus características. Utilizaremos entonces muestras apareadas.

Para ello usamos la opción *Análisis de datos* del menú *Herramientas,* y en la lista *Funciones para análisis* elegimos *Prueba t para medias de dos muestras emparejadas* (Figura 10-62). Rellenamos la pantalla de entrada tal y como se indica en la Figura 10-63, y al pulsar *Aceptar* se obtienen los resultados de la Figura 10-64.

Según los resultados, se acepta que las calificaciones de los dos cursos son idénticas en media al 95% de confianza ya que los p-valores para los contrastes unilateral y bilateral son mayores que 0,05. Además, ambos valores críticos son mayores que el valor del estadístico *T*.

CAPÍTULO XI

REDUCCIÓN DE LA DIMENSIÓN MEDIANTE ANÁLISIS DE COMPONENTES PRINCIPALES Y ANÁLISIS FACTORIAL

CÉSAR PÉREZ LÓPEZ
DANIEL SANTÍN GONZÁLEZ

11.1. INTRODUCCIÓN A LAS TÉCNICAS DE REDUCCIÓN DE LA DIMENSIÓN

Es habitual en el trabajo estadístico específico de la investigación social disponer de muchas variables medidas u observadas en una colección de individuos y pretender estudiarlas conjuntamente, para lo cual se suele acudir al análisis estadístico multivariante de datos. Entonces se dispone de una diversidad de técnicas y debe seleccionarse la más adecuada a los datos y al objetivo científico. Al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los *métodos multivariantes de reducción de la dimensión* (análisis en componentes principales, factorial, correspondencias, escalamiento óptimo y multidimensional, etc.) tratan de eliminarla. Estos métodos combinan muchas variables observadas para obtener pocas variables ficticias que las representen con la mínima pérdida de información.

Estos métodos de reducción de la dimensión son *métodos multivariantes de la interdepedencia* en el sentido de que todas sus variables tienen una importancia equivalente, es decir, si ninguna variable destaca como dependiente principal en el objetivo de la investigación. En este caso también deberá tener en cuenta el tipo de variables que se maneja. Si son variables cuantitativas, las técnicas de reducción de la dimensión pueden ser el *Análisis de Componentes Principales* y el *Análisis Factorial*, si son variables cualitativas, puede acudirse al *Análisis de Correspondencias* (Tema 13) y si son variables cualitativas ordinales se acude al *Escalamiento Multidimensional* (Tema 14).

Los métodos de la interdependencia se contraponen a los denominados *métodos multivariantes de la dependencia* en los cuales no es aceptable una importancia equivalente en las variables, porque alguna se destaca como dependiente principal. En este caso habrá de utilizar técnicas multivariantes analíticas o inferenciales considerando la variable dependiente como explicada por las demás variables independientes explicativas, y tratando de relacionar todas las variables por medio de una posible ecuación o modelo que las li-

gue. El método elegido podría ser entonces la Regresión Lineal (Tema 15), generalmente con todas las variables cuantitativas. Una vez configurado el modelo matemático se podrá llegar a predecir el valor de la variable dependiente conocido el perfil de todas las demás. Si la variable dependiente fuera cualitativa dicotómica (1,0; sí o no) podrá usarse como clasificadora, estudiando su relación con el resto de variables clasificadoras a través de la Regresión Logística (Tema 16). Si la variable dependiente cualitativa observada constatara la asignación de cada individuo a grupos previamente definidos (dos, o más de dos), puede ser utilizada para clasificar nuevos casos en los que se desconozca el grupo al que probablemente pertenecen, en cuyo caso estamos ante el Análisis Discriminante, que resuelve el problema de asignación en función de un perfil cuantitativo de variables clasificativas. Si la variable dependiente es cuantitativa y las explicativas son cualitativas estamos ante los modelos del análisis de la varianza, que puede extenderse a los modelos loglineales para el análisis de tablas de contingencia de dimensión elevada. Si la variable dependiente puede ser cualitativa o cuantitativa y las independientes cualitativas, estamos ante la Segmentación.

11.2. ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis en componentes principales es una técnica de análisis estadístico multivariante que se clasifica entre los métodos de interdependencia. Se trata de un método multivariante de simplificación o reducción de la dimensión y que se aplica cuando se dispone de un conjunto elevado de variables con datos cuantitativos correlacionadas entre sí persiguiendo obtener un menor número de variables, combinación lineal de las primitivas e incorrelacionadas, que se denominan componentes principales o factores, que resuman lo mejor posible a las variables iniciales con la mínima pérdida de información y cuya posterior interpretación permitirá un análisis más simple del problema estudiado. Esta reducción de muchas variables a pocas componentes puede simplificar la aplicación sobre estas últimas de otras técnicas multivariantes (regresión, clusters, etc.) y permite la obtención de indicadores robustos para ordenar individuos o familias de acuerdo a sus características.

El elevado número de variables iniciales x_1, x_2,…,x_p se resumen en unas pocas variables C_1, C_2,…,C_k (*componentes principales*) *perfectamente calculables* y que sintetizan la mayor parte de la información contenida en sus datos. Inicialmente se tienen tantas componentes como variables:

$$C_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$
$$\vdots$$
$$C_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p$$

Pero sólo se retienen las k componentes principales que explican un porcentaje alto de la variabilidad de las variables iniciales (C_1, C_2,…, C_k).

Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza mayor es la información que lleva incorporada dicha componente. Por esta razón se selecciona como primera componente aquella que tenga mayor varianza, mientras que, por el contrario, la última es la de menor varianza.

En general, la extracción de componentes principales se efectúa sobre variables *tipificadas* para evitar problemas derivados de escala, aunque también se puede aplicar sobre variables expresadas en *desviaciones* respecto a la media.

Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes. Si las variables originales estuvieran completamente incorrelacionadas entre sí, entonces el análisis de componentes principales carecería por completo de interés, ya que en ese caso las componentes principales coincidirían con las variables originales.

Cálculo de las componentes principales

En el análisis en componentes principales se dispone de una muestra de tamaño n acerca de p variables X_1, X_2,...,X_p (tipificadas o expresadas en desviaciones respecto de su media) inicialmente correlacionadas, para posteriormente obtener a partir de ellas un número $k \leq p$ de variables incorrelacionadas C_1, C_2,...,C_k que sean combinación lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad. *La primera componente principal*, al igual que las restantes, se expresa como combinación lineal de las variables originales como sigue:

$$C_i = u_{i1}X_1 + u_{i2}X_2 + \cdots + u_{ip}X_p \quad i = 1,\ldots,n$$

Para el conjunto de las n observaciones muestrales y para todas las componentes tenemos:

$$\begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} \cdots X_{p1} \\ X_{12} & X_{22} \cdots X_{p2} \\ & \vdots \\ X_{1n} & X_{2n} \cdots X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

En notación abreviada tendremos: $C_1 = X\,u_1$ y:

$$V(C_1) = \frac{\sum_{i=1}^{n} C_{1i}^2}{n} = \frac{1}{n} C_1' C_1 = \frac{1}{n} u_1' X' X u_1 = u_1' \left[\frac{1}{n} X' X\right] u_1 = u_1' V u_1$$

La primera componente C_1 se obtiene de forma que su varianza sea máxima sujeta a la restricción de que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad, es decir, la variable de los pesos o ponderaciones $(u_{11}, u_{12},...,u_{1p})'$ se toma normalizada. Se trata entonces de hallar C_1 maximizando $V(C_1) = u_1'Vu_1$, sujeta a la restricción:

$$\sum_{j=1}^{p} u_{1j}^2 = u_1' u_1 = 1$$

Se demuestra que, para maximizar $V(C_1)$ se toma el mayor valor propio λ_1 de la matriz V. Sea λ_1 el citado mayor valor propio de V y tomando u_1 como su vector propio asociado normalizado ($u_1'u_1$=1), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la primera componente principal, componente que vendrá definida como:

$$C_1 = u_1 X = u_{11} X_1 + u_{12} X_2 + \cdots + u_{1p} X_p$$

Para maximizar $V(C_2)$ hemos de tomar el segundo mayor valor propio λ_2 de la matriz V (el mayor ya lo había tomado al obtener la primera componente principal).

Tomando λ_2 como el segundo mayor valor propio de V y tomando u_2 como su vector propio asociado normalizado ($u_2'u_2$=1), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la segunda componente principal, componente que vendrá definida como:

$$C_2 = u_2 X = u_{21} X_1 + u_{22} X_2 + \cdots + u_{2p} X_p$$

De forma similar, la componente principal h-ésima se define como $C_h = Xu_h$ donde u_h es el vector propio de V asociado a su h-ésimo mayor valor propio. Suele denominarse también a u_h eje factorial h-ésimo.

Se demuestra que la proporción de la variabilidad total recogida por la componente principal h-ésima (*porcentaje de inercia explicada por la componente principal h-ésima*) vendrá dada por:

$$\frac{\lambda_h}{\sum_{h=1}^{p}\lambda_h} = \frac{\lambda_h}{traza(V)}$$

Si las variables están tipificadas, *traza(V) = p*, con lo que la proporción de la componente h-esima en la variabilidad total será λ_h/p. También se define el *porcentaje de inercia explicada por las k primeras componentes principales (o ejes factoriales)* como:

$$\frac{\sum_{h=1}^{k}\lambda_h}{\sum_{h=1}^{p}\lambda_h} = \frac{\sum_{h=1}^{k}\lambda_h}{traza(V)}$$

Puntuaciones o medición de las componentes

El análisis en componentes principales es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por las componentes obtenidas. Por ejemplo en el caso de estimación de modelos afectados de multicolinealidad o correlación serial (autocorrelación). Por ello, es necesario conocer los valores que toman las componentes en cada observación.

Una vez calculados los coeficientes u_{hj} (componentes del vector propio normalizado asociado al valor propio h-ésimo de la matriz $V = X'X/n$ relativo a la componente principal Z_h), se pueden obtener las puntuaciones Z_{hj}, es decir, los valores de las componentes correspondientes a cada observación, a partir de la siguiente relación:

$$Z_{hi} = u_{h1}X_{1i} + u_{h2}X_{2i} + \cdots + u_{hp}X_{pi} \quad h = 1,\ldots,p \quad i = 1,\ldots,n$$

Número de componentes principales a retener

En general, el objetivo de la aplicación de las componentes principales es reducir las dimensiones de las variables originales, pasando de *p* variables originales a *m<p* componentes principales. El problema que se plantea es cómo fijar *m*, o, dicho de otra forma, ¿qué número de componentes se deben retener? Aunque para la extracción de las componentes principales no hace falta plantear un modelo estadístico previo, algunos de los criterios para determinar cuál debe ser el número óptimo de componentes a retener requieren la formulación previa de hipótesis estadísticas.

Criterio de la media aritmética

Según este criterio se seleccionan aquellas componentes cuya raíz característica λ_j excede de la media de las raíces características. Recordemos que la raíz característica asociada a una componente es precisamente su varianza. Analíticamente este criterio implica retener todas aquellas componentes en que se verifique que:

$$\lambda_j > \bar{\lambda} = \frac{\sum_{j=1}^{p} \lambda_j}{p}$$

Si se utilizan variables tipificadas, entonces, como ya se ha visto, se verifica que $\sum_{j=1}^{p} \lambda_j = p$, con lo que para variables tipificadas se retiene aquellas componentes tales que $\lambda_j > 1$.

Criterio del gráfico de sedimentación

El *gráfico de sedimentación* se obtiene al representar en ordenadas las raíces características y en abscisas los números de las componentes principales correspondientes a cada raíz característica en orden decreciente. Uniendo todos los puntos se obtiene una Figura que, en general, se parece al perfil de una montaña con una pendiente fuerte hasta llegar a la base, formada por una meseta con una ligera inclinación. Continuando con el símil de la montaña, en esa meseta es donde se acumulan los guijarros caídos desde la cumbre, es decir, donde se sedimentan. Por esta razón, a este gráfico se le conoce con el nombre de gráfico de sedimentación. Su denominación en inglés es *scree plot.* De acuerdo con el criterio gráfico se retienen todas aquellas componentes previas a la zona de sedimentación.

Matriz de cargas factoriales, comunalidad y círculos de correlación

La dificultad en la interpretación de los componentes estriba en la necesidad de que tengan sentido y midan algo útil en el contexto del fenómeno estudiado. Por tanto, es indispensable considerar el peso que cada variable original tiene dentro del componente elegido, así como las correlaciones existentes entre variables y factores. Un componente es una función lineal de todas las variables, pero puede estar muy bien correlacionado con algunas de ellas, y menos con otras. Ya hemos visto que el coeficiente de correlación entre una componente y una variable se calcula

multiplicando el peso de la variable en esa componente por la raíz cuadrada de su valor propio:

$$r_{ik} = u_{ik}\sqrt{\lambda_k}$$

Se demuestra también que estos coeficientes *r* representan la parte de varianza de cada variable que explica cada factor. De este modo, cada variable puede ser representada como una función lineal de los *k* componentes retenidos, donde los pesos o cargas de cada componente o factor (*cargas factoriales*) en la variable coinciden con los coeficientes de correlación.

El cálculo matricial permite obtener de forma inmediata la tabla de coeficientes de correlación variables-componentes (*pxk*), que se denomina *matriz de cargas factoriales*. Las ecuaciones de las variables en función de las componentes (factores), traspuestas las inicialmente planteadas, son de mayor utilidad en la interpretación de los componentes, y se expresan como sigue:

$$\begin{aligned} Z_1 &= r_{11}X_1 + \cdots + r_{1p}X_p \\ Z_2 &= r_{21}X_1 + \cdots + r_{2p}X_p \\ &\vdots \\ Z_k &= r_{k1}X_1 + \cdots + r_{kp}X_p \end{aligned} \quad \Rightarrow \quad \begin{aligned} X_1 &= r_{11}Z_1 + \cdots + r_{k1}Z_k \\ X_2 &= r_{12}Z_1 + \cdots + r_{k2}Z_k \\ &\vdots \\ X_p &= r_{1p}Z_1 + \cdots + r_{kp}Z_k \end{aligned}$$

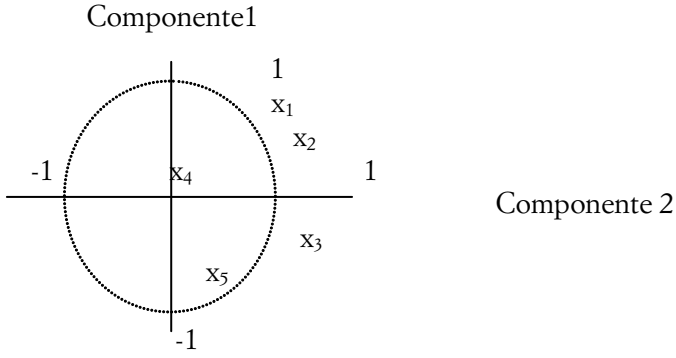
Para la primera variable, la comunalidad será $r^2_{11} + ... + r^2_{k1} = V(X_1) = h_1^2$. Por consiguiente, la suma de las comunalidades de todas las variables representa la parte de inercia global de la nube original explicada por los *k* factores retenidos, y coincide con la suma de los valores propios de estas componentes.

La comunalidad proporciona un criterio de calidad de la representación de cada variable, de modo que, variables totalmente representadas tienen de comunalidad la unidad.

También se demuestra que la suma en vertical de los cuadrados de las cargas factoriales de todas las variables en un componente es su valor propio. Por ejemplo, el valor propio del primer componente será $r^2_{11} + ... + r^2_{1p} = 1$.

Es evidente que, al ser las cargas factoriales los coeficientes de correlación entre variables y componentes, su empleo hace comparables los pesos de cada variable en la componente y facilita su interpretación. En este mismo sentido, su representación gráfica puede orientar al investigador en una primera aproximación a la interpretación de los componentes. Como es lógico, esta representación sobre un plano sólo puede contener los fac-

tores de dos en dos, por lo que se pueden realizar tantos gráficos como parejas de factores retenidos. Estos gráficos se denominan *círculos de correlación,* y están formados por puntos que representan cada variable por medio de dos coordenadas que miden los coeficientes de correlación de dicha variable con los dos factores o componentes considerados. Todas las variables estarán contenidas dentro de un círculo de radio unidad.



Rotación de las componentes

Es frecuente no encontrar interpretaciones verosímiles a los factores (componentes) obtenidos. Sería deseable, para una más fácil interpretación, que cada componente estuviera relacionada muy bien con pocas variables (coeficientes de correlación *r* próximos a 1 ó -1) y mal con las demás (*r* próximos a 0). Esta optimización se obtiene por una adecuada *rotación de los ejes* que definen los componentes principales.

Rotar un conjunto de componentes no cambia la proporción de inercia total explicada, como tampoco cambia las comunalidades de cada variable, que no son sino la proporción de varianza explicada por todos ellos. Las rotaciones más utilizadas entre las muchas existentes son la rotación VARIMAX y la QUARTIMAX (ortogonales) y PROMAX (oblicua).

Sin embargo, los coeficientes, que dependen directamente de la posición de los componentes respecto a las variables originales (cargas factoriales y valores propios), se ven alterados por la rotación.

11.3. Análisis factorial

El *análisis factorial* tiene como objeto simplificar las múltiples y complejas relaciones que puedan existir entre un conjunto de variables obser-

vadas X_1, X_2,...,X_p. Para ello trata de encontrar dimensiones comunes o *factores* que ligan a las aparentemente no relacionadas variables. Concretamente, se trata de encontrar un conjunto de $k<p$ *factores no directamente observables* F_1, F_2,...,F_k que expliquen suficientemente a las variables observadas perdiendo el mínimo de información, de modo que sean fácilmente interpretables (*principio de interpretabilidad*) y que sean los menos posibles, es decir, k pequeño (*principio de parsimonia*). Además, los factores han de extraerse de forma que resulten independientes entre sí, es decir, que sean ortogonales. En consecuencia, el análisis factorial es una técnica de reducción de datos que examina la interdependencia de variables y proporciona conocimiento de la estructura subyacente de los datos.

El aspecto más característico del análisis factorial lo constituye su capacidad de reducción de datos. Las relaciones entre las variables observadas X_1 X_2,...,X_p vienen dadas por su matriz de correlaciones, cuyo determinante ha de ser pequeño (hay relación entre ellas).

El análisis de componentes principales y el análisis factorial tienen en común que son técnicas de reducción de la dimensión para examinar la interdependencia de variables, pero difieren en su objetivo, sus características y su grado de formalización. La diferencia entre análisis en componentes principales y análisis factorial radica en que en el análisis factorial trata de encontrar variables sintéticas latentes, inobservables y aún no medidas cuya existencia se sospecha en las variables originales y que permanecen a la espera de ser halladas, mientras que en el análisis en componentes principales se obtienen variables sintéticas combinación de las originales y cuyo cálculo es posible basándose en aspectos matemáticos independientes de su interpretabilidad práctica.

En el análisis en componentes principales la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan (sus componentes). Pero esto no ocurre en el análisis factorial.

En el análisis factorial sólo una parte de la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan (*factores comunes* F_1, F_2,…,F_p). Esta parte de la variabilidad de cada variable original explicada por los factores comunes se denomina *comunalidad*, mientras que la parte de varianza no explicada por los factores comunes se denomina *unicidad* (*comunalidad* + *unicidad* = 1) y representa la parte de variabilidad propia f_i de cada variable x_i.

$$x_1 = c_{11}F_1 + c_{12}x_2 + \cdots + c_{1p}F_p + f_1$$
$$\vdots$$
$$x_n = c_{n1}F_1 + c_{n2}F_2 + \cdots + c_{np}F_p + f_n$$

Cuando la comunalidad es unitaria (*unicidad nula*) el análisis en componentes principales coincide con el factorial. Es decir, el análisis en componentes principales es un caso particular del análisis factorial en el que los factores comunes explican el 110% de la varianza total.

Entre los métodos para obtener los factores destacan los siguientes:

- *Método de las componentes principales (presente en XLSTAT).*
- *Métodos MINRES (minimización residual), ULS (mínimos cuadrados no ponderados) y GLS (mínimos cuadrados generalizados).*
- *Método de máxima verosimilitud (presente en XLSTAT).*
- *Método de componentes principales iteradas o ejes principales.*
- *Método del factor principal (presente en XLSTAT).*
- *Método alfa.*
- *Método de factorización imagen.*
- *Método del centroide.*
- *Método de Turstone.*

A continuación se presentan las características de los métodos más importantes de extracción de los factores.

- *Método de las componentes principales*. Método de extracción de factores utilizado para formar combinaciones lineales no correlacionadas de las variables observadas. La primera componente tiene la varianza máxima. Las componentes sucesivas explican progresivamente proporciones menores de la varianza y no están correlacionadas las unas con las otras. El análisis de componentes principales se utiliza para obtener la solución factorial inicial. Puede utilizarse cuando una matriz de correlaciones es singular.
- *Método de mínimos cuadrados no ponderados*. Método de extracción factorial que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlaciones observada y reproducida, ignorando las diagonales.
- *Método de mínimos cuadrados generalizados*. Método de extracción de factores que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlación observada y reproducida. Las correlaciones se ponderan por el inverso de su unicidad, de manera que las variables que tengan un valor alto de unicidad reciban un peso menor que las que tengan un valor bajo de unicidad.

- *Método de máxima verosimilitud*. Método de extracción factorial que proporciona las estimaciones de los parámetros que con mayor probabilidad han producido la matriz de correlaciones observada, si la muestra procede de una distribución normal multivariada. Las correlaciones se ponderan por el inverso de la unicidad de las variables y se emplea un algoritmo iterativo.
- *Factorización de ejes principales*. Método de extracción de factores que parte de la matriz de correlaciones original con los cuadrados de los coeficientes de correlación múltiple insertados en la diagonal principal como estimaciones iniciales de las comunalidades. Las saturaciones factoriales resultantes se utilizan para estimar de nuevo las comunalidades y reemplazan a las estimaciones previas en la diagonal de la matriz. Las iteraciones continúan hasta que los cambios en las comunalidades, de una iteración a la siguiente, satisfagan el criterio de convergencia para la extracción.
- *Alfa*. Método de extracción factorial que considera a las variables incluidas en el análisis como una muestra del universo de las variables posibles. Este método maximiza el Alfa de Cronbach para los factores.
- *Factorización imagen*. Método de extracción de factores, desarrollado por Guttman y basado en la teoría de las imágenes. La parte común de una variable, llamada la imagen parcial, se define como su regresión lineal sobre las restantes variables, en lugar de ser una función de los factores hipotéticos.

Contrastes en el modelo factorial

En el modelo factorial pueden realizarse varios tipos de contrastes. Estos contrastes suelen agruparse en dos bloques, según se apliquen previamente a la extracción de los factores o se apliquen después.

Con los contrastes aplicados previamente a la extracción de los factores trata de analizarse la pertinencia de la aplicación del análisis factorial a un conjunto de variables observables. Con los contrastes aplicados después de la obtención de los factores se pretende evaluar el modelo factorial una vez estimado.

Dentro del grupo de *contrastes que se aplican previamente a la extracción de los factores* tenemos el contraste de esfericidad de Barlett y la medida de adecuación muestral de Kaiser, Meyer y Olkin.

Evidentemente, antes de realizar un análisis factorial nos plantearemos si las p variables originales están correlacionadas entre sí o no lo están. Si no lo estuvieran no existirían factores comunes y, por lo tanto, no

tendría sentido aplicar el análisis factorial. Esta cuestión suele probarse utilizando el contraste de esfericidad de Barlett que se basa en que la matriz de correlación poblacional R_p recoge la relación entre cada par de variables mediante sus elementos r_{ij} situados fuera de la diagonal principal. Los elementos de la diagonal principal son unos, ya que toda variable está totalmente relacionada consigo misma. En caso de que no existiese ninguna relación entre las p variables en estudio, la matriz R_p sería la identidad, cuyo determinante es la unidad. Por lo tanto, para decidir la ausencia o no de relación entre las p variables puede plantearse el siguiente contraste:

$$H_0 : |R_p| = 1$$

$$H_1 : |R_p| \cdot 1$$

Barlett introdujo un estadístico para este contraste basado en la matriz de correlación muestral R, que bajo la hipótesis H_0 tiene una distribución *Chi-cuadrado* con $p(p\text{-}1)/2$ grados de libertad. La expresión de este estadístico es la siguiente:

$$-[n\text{-}11\text{-}(2p+5)/6]Ln|R|$$

Por otro lado, Kaiser-Meyer y Olkin definen la medida *KMO* de adecuación muestral global al modelo factorial basada en los coeficientes de correlación observados de cada par de variables y en sus coeficientes de correlación parcial mediante la expresión siguiente:

$$KMO = \frac{\sum_j \sum_{h \neq j} r_{jh}^2}{\sum_j \sum_{h \neq j} r_{jh}^2 + \sum_j \sum_{h \neq j} a_{jh}^2}$$

r_{jh} son los coeficientes de correlación observados entre las variables X_j y X_h

a_{jh} son los coeficientes de correlación parcial entre las variables X_j y X_h

En el caso de que exista adecuación de los datos a un modelo de análisis factorial, el término del denominador, que recoge los coeficientes a_{jh}, será pequeño y, en consecuencia, la medida *KMO* será próxima a la unidad. Valores de *KMO* por debajo de 0,5 no serán aceptables, considerándose inadecuados los datos a un modelo de análisis factorial. Para valores superiores a 0,5 se considera aceptable la adecuación de los datos a un modelo de análisis factorial. Mientras más cercas estén de 1 los valores de

KMO mejor es la adecuación de los datos a un modelo factorial, considerándose ya excelente la adecuación para valores de *KMO* próximos a 0,9.

También existe una medida de adecuación muestral individual para cada una de las variables basada en la medida *KMO*. Esta medida se denomina *MSA* (*Measure of Sampling Adequacy*), se define de la siguiente forma:

$$MSA_j = \frac{\sum_{h \neq j} r_{jh}^2}{\sum_{h \neq j} r_{jh}^2 + \sum_{h \neq j} a_{jh}^2}$$

Si el valor de MSA_j se aproxima a la unidad, la variable X_j será adecuada para su tratamiento en el análisis factorial con el resto de las variables.

También en el modelo factorial pueden realizarse *contrastes después de la obtención de los factores con los que se pretende evaluar el modelo factorial una vez estimado*. Entre ellos tenemos el contraste para la bondad de ajuste del método de máxima verosimilitud y el contraste para la bondad de ajuste del método MINRES.

Rotación de los factores

El trabajo en el análisis factorial persigue que los factores comunes tengan una interpretación clara, porque de esa forma se analizan mejor las interrelaciones existentes entre las variables originales. Sin embargo, en muy pocas ocasiones resulta fácil encontrar una interpretación adecuada de los factores, iniciales, con independencia del método que se haya utilizado para su extracción. Precisamente los procedimientos de `rotaci n de factores` se han ideado para obtener, a partir de la solución inicial, unos factores que sean fácilmente interpretables.

Rotaciones ortogonales

- *Método Varimax.*
- *Método Quartimax.*
- *Métodos Ortomax: Ortomax general, Biquartimax y Equamax.*

A continuación se presentan las características de los métodos más importantes de rotación ortogonal.

- *Método varimax*. Método de rotación ortogonal que minimiza el número de variables que tienen saturaciones altas en cada factor. Simplifica la interpretación de los factores.

- *Método quartimax*. Método de rotación que minimiza el número de factores necesarios para explicar cada variable. Simplifica la interpretación de las variables observadas.

- *Método equamax*. Método de rotación que es combinación del método varimax, que simplifica los factores, y el método quartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.

Rotaciones oblicuas

- *Método Oblimax y método Quartimin.*
- *Métodos Oblimin: Covarimin, Oblimin directo (o general) y Biquartimin.*
- *Método Oblimin directo: Rotación Promax.*

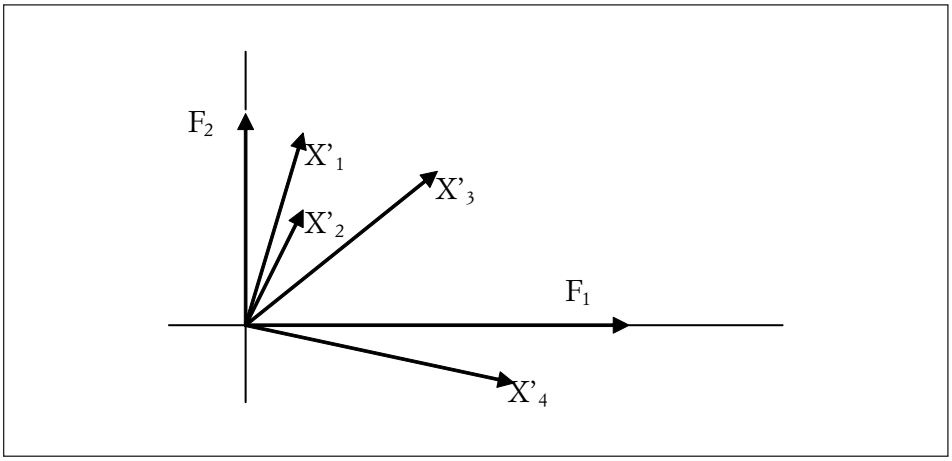
A continuación se presentan las características de los métodos más importantes de rotación oblícua.

- *Criterio Oblimin directo*. Método para la rotación oblicua (no ortogonal). Cuando delta es igual a cero (el valor por defecto) las soluciones son las más oblicuas. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el valor por defecto 0 para delta, introduzca un número menor o igual que 0,8.

- *Rotación promax*. Rotación oblicua que permite que los factores estén correlacionados. Puede calcularse más rápidamente que una rotación oblimin directa, por lo que es útil para conjuntos de datos grandes.

Interpretación gráfica de los factores

A continuación se presenta un gráfico relativo a cuatro variables X_1, X_2, X_3 y X_4 representadas por dos factores F_1 y F_2.



Como las saturaciones, pesos o cargas factoriales de cada variable en cada factor (elementos de la matriz factorial), se representan por las proyecciones ortogonales de cada variable en cada factor, la cuarta variable se explica fuertemente y de forma positiva por el primer factor (proyección positiva grande de X'_4 sobre F_1), mientras que se representa poco y en sentido negativo por el segundo factor (proyección negativa pequeña de X'_4 sobre F_2). De la misma forma, la primera y segunda variables se explican fuertemente y de forma positiva por el segundo factor, y se explican poco y de forma positiva por el primer factor. La tercera variable se explica de igual forma por el primero y segundo factor.

Si la representación geométrica resulta difusa, se puede realizar una rotación de los factores que clarifique las proyecciones de las variables sobre ellos. Con una rotación factorial se transforma una solución factorial inicial en otro tipo de solución preferida. Tal transformación va encaminada a poner de manifiesto la solución de la manera más convincente y clara para su interpretación científica.

Puntuaciones o medición de los factores

El análisis factorial es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por los factores obtenidos. Por ejemplo en el caso de estimación de modelos afectados de multicolinealidad. Por ello, es necesario conocer los valores que toman los factores en cada observación (puntuaciones factoriales). Sin embargo, es importante hacer constar que, salvo el caso de que se haya aplicado el análisis de componentes principales para la extracción de factores, no se obtienen unas puntuaciones exactas para los factores.

En su lugar, es preciso realizar estimaciones para obtenerlas. Estas estimaciones se pueden realizar por distintos métodos. Los procedimientos más conocidos, y que aparecen implementados en los paquetes de software son los de *mínimos cuadrados, regresión, Anderson-Rubin y Barlett.*

Las características de los métodos más importantes de obtención de las componentes se presentan a continuación.

- *Método de regresión*. Método de estimación de los coeficientes de las puntuaciones factoriales. Las puntuaciones resultantes tienen de media 0 y varianza igual al cuadrado de la correlación múltiple entre las puntuaciones factoriales estimadas y los valores factoriales verdaderos. Las puntuaciones pueden estar correlacionadas incluso cuando los factores son ortogonales.

- *Puntuaciones de Bartlett*. Método de estimación de los coeficientes para las puntuaciones factoriales. Las puntuaciones resultantes tienen una media de 0. Se minimiza la suma de cuadrados de los factores únicos sobre el rango de las variables.

- *Método de Anderson-Rubin*. Método de estimación de los coeficientes para las puntuaciones factoriales. Es una modificación del método de Bartlett, que asegura la ortogonalidad de los factores estimados. Las puntuaciones resultantes tienen una media 0, una desviación típica de 1 y no están correlacionadas.

EJERCICIO 11-1. El fichero CAF contiene información referente a 7 indicadores de resultados y de actuaciones obtenidos por los 14 centros de atención a la familia (CAF) de una región.

Se pide realizar un análisis de componentes principales con el objetivo de reducir la información a un número reducido de indicadores que sinteticen estos datos.

Para ejecutar este análisis utilizaremos XLSTAT → Análisis de Datos → Análisis de Componentes Principales. En la pestaña General podemos introducir los datos en forma de variables o bien introducir directamente la matriz de correlaciones o covarianzas.

FIGURA 11-1

Una vez dada la información acerca de los datos, la pestaña Resultados permite especificar los resultados que queremos obtener (Figura 11-2).

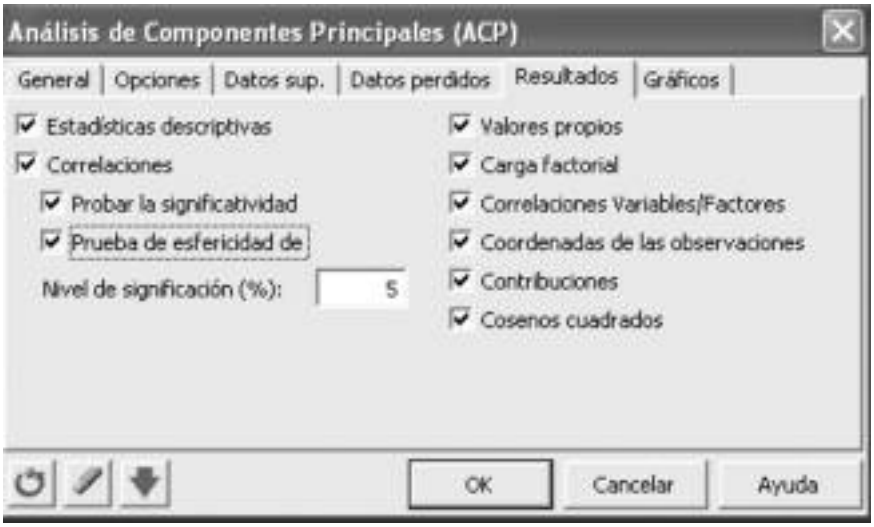


FIGURA 11-2

Una vez especificadas las opciones obtenemos los resultados del análisis cuyos resultados se detallan.

Matriz de correlaciones

Variables	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
X1	1	-0,150	0,019	0,490	-0,131	-0,255	-0,069
X2	-0,150	**1**	**-0,786**	-0,183	**0,890**	**-0,715**	**-0,783**
X3	0,019	**-0,786**	**1**	0,196	**-0,602**	**0,830**	**0,722**
X4	0,490	-0,183	0,196	1	0,002	0,009	0,134
X5	-0,131	**0,890**	**-0,602**	0,002	**1**	-0,494	-0,526
X6	-0,255	**-0,715**	**0,830**	0,009	-0,494	**1**	**0,915**
X7	-0,069	**-0,783**	**0,722**	0,134	-0,526	**0,915**	**1**

Los valores en negrita son significativamente diferentes de 0 con un nivel de significación alfa=0,05

Prueba de esfericidad de Barlett

Chi-cuadrado ajustado (Valor observado)	77,787
Chi-cuadrado ajustado (Valor crítico)	32,671
GDL	21
p-valor	< 0,0001
alfa	0,05

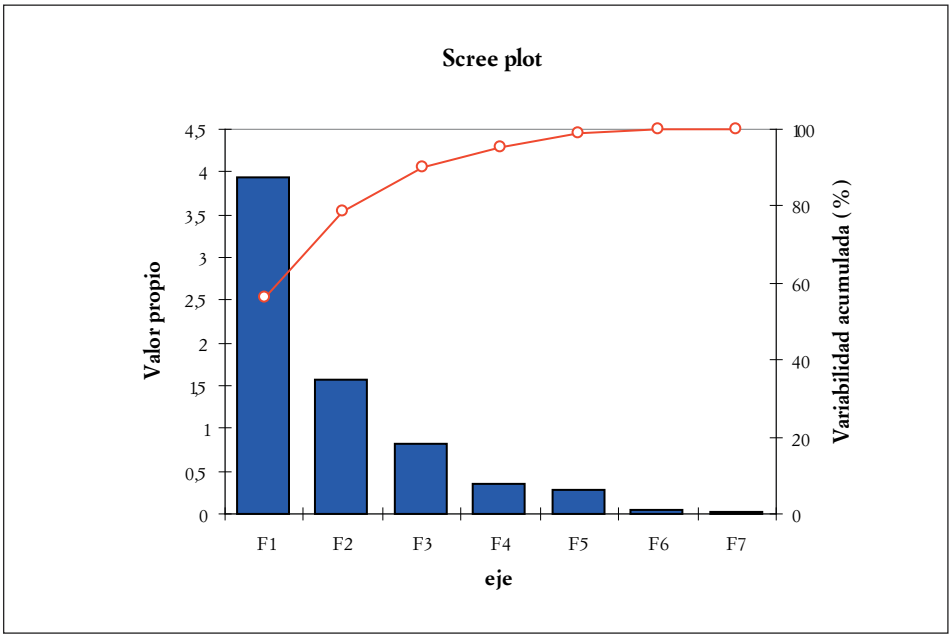
Como el p-valor computado en la prueba de Barlett es menor que el nivel de significación alfa=0,05, se debe rechazar la hipótesis nula H_0 de que los coeficientes de correlación sean iguales a cero, y aceptar la hipótesis alternativa H_a.

Valores Propios* (autovalores)

	Valor propio	*Variabilidad (%)*	*% acumulado*
F1	3,937	56,238	56,238
F2	1,564	22,342	78,580
F3	0,810	11,571	90,151
F4	0,357	5,102	95,253
F5	0,270	3,861	99,114
F6	0,045	0,644	99,759
F7	0,017	0,241	100,000

** Se ha transpuesto la matriz de resultados.*

Los valores propios muestran como las dos primeras componentes explican el 78,58% de la varianza total de los datos (Suma de F1 y F2). Hay dos valores propios mayores que uno, lo que induce a tomar las dos primeras componentes como resumen de las 7 variables.



Vectores Propios (Autovectores)

Variables	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
X1	-0,009	0,717	-0,252	0,634	0,060	0,116	-0,059
X2	0,476	-0,108	0,257	0,156	0,158	0,598	0,537
X3	-0,452	0,015	0,146	0,022	0,805	-0,189	0,301
X4	-0,083	0,639	0,538	-0,519	-0,116	0,114	0,003
X5	0,394	-0,067	0,626	0,379	0,097	-0,462	-0,286
X6	-0,450	-0,234	0,311	0,227	0,016	0,577	-0,511
X7	-0,452	-0,085	0,267	0,329	-0,549	-0,182	0,524

Las componentes principales se calculan a partir de la combinación lineal entre el valor de los autovectores y las variables originales estandarizadas.

$$C1 = -0.009*X1 + 0,476*X2 - 0,452*X3 - 0,083*X4 + 0,394*X5 - 0,450*X6 - 0,452*X7$$

De esta forma es como se obtiene la matriz de coordenadas de las observaciones que se muestra posteriormente. También se puede establecer a partir de las cargas factoriales la relación entre las variables originales estandarizadas y el valor calculado.

X1 = -0,018F1 + 0,896 F2 -0,227F3 + 0,379F4 + 0,031F5 + 0,025F6 -0,008F7

Cargas factoriales (correlación entre variables y factores)

	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
X1	-0,018	0,896	-0,227	0,379	0,031	0,025	-0,008
X2	0,944	-0,135	0,231	0,093	0,082	0,127	0,070
X3	-0,897	0,018	0,131	0,013	0,418	-0,040	0,039
X4	-0,165	0,799	0,484	-0,310	-0,060	0,024	0,000
X5	0,781	-0,084	0,564	0,226	0,051	-0,098	-0,037
X6	-0,893	-0,293	0,280	0,136	0,008	0,123	-0,066
X7	-0,897	-0,106	0,240	0,197	-0,285	-0,039	0,068

Observando las cargas factoriales se puede deducir que la primera componente está correlacionada con las variables X2, X3, X5, X5 y X7 mientras que la segunda lo está con X1 y X4.

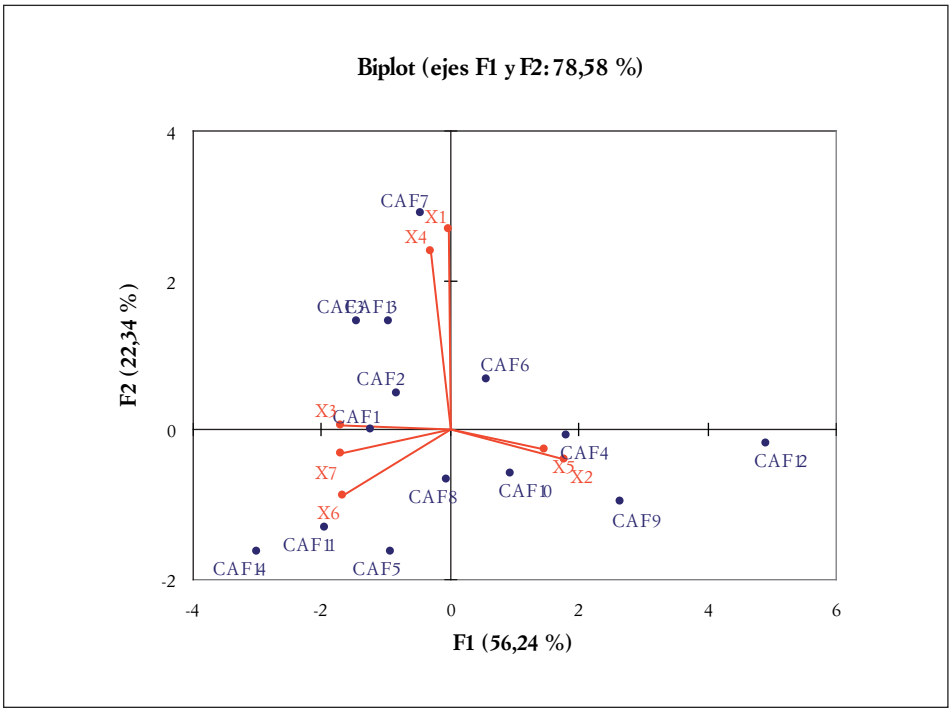
Contribuciones de las variables (%)

	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
X1	0,008	51,341	6,340	40,250	0,358	1,354	0,349
X2	22,651	1,173	6,596	2,447	2,503	35,787	28,842
X3	20,429	0,021	2,120	0,047	64,740	3,576	9,067
X4	0,690	40,803	28,932	26,928	1,340	1,307	0,001
X5	15,511	0,446	39,215	14,335	0,946	21,368	8,178
X6	20,265	5,491	9,691	5,146	0,024	33,303	26,080
X7	20,446	0,725	7,107	10,847	30,089	3,305	27,481
Total	100	100	100	100	100	100	100

Coordenadas de las observaciones

Observación	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
CAF1	-1,240	0,005	1,006	-1,220	-0,531	0,024	-0,133
CAF2	-0,844	0,500	0,156	-0,596	0,464	-0,218	0,185
CAF3	-1,466	1,445	0,065	0,662	0,747	-0,143	0,069
CAF4	1,811	-0,065	-0,547	-0,666	-0,113	0,443	0,212
CAF5	-0,930	-1,618	-0,245	-0,049	0,867	0,195	-0,220
CAF6	0,564	0,682	-0,889	0,293	-0,223	-0,016	-0,071
CAF7	-0,449	2,891	0,035	0,018	0,164	0,332	-0,082
CAF8	-0,057	-0,653	-0,242	-0,758	-0,292	-0,272	-0,102
CAF9	2,661	-0,948	-1,515	0,580	-0,171	0,004	-0,115
CAF10	0,949	-0,585	-1,016	-0,124	-0,225	-0,177	0,186
CAF11	-1,960	-1,292	0,134	0,069	0,688	-0,082	0,044
CAF12	4,927	-0,183	2,192	0,388	0,184	-0,082	-0,008
CAF13	-0,965	1,447	-0,101	0,455	-0,759	-0,228	-0,065
CAF14	-3,001	-1,626	0,966	0,948	-0,800	0,220	0,102

Representación gráfica de los dos primeros factores



Los vectores señalan la dirección para cada variable. Todas las variables (salvo X6) están únicamente correlacionadas fundamentalmente con uno de los ejes. Este hecho hace que en este caso no sea necesaria una rotación de los ejes para que factores y variables queden correlacionados entre sí de forma que no se solapen. Conviene recalcar que al ser los ejes obtenidos ortogonales no están correlacionados por lo que estas puntuaciones pueden utilizarse para otros fines como regresión múltiple.

EJERCICIO 11-2. Repita el ejercicio anterior utilizando el método de máxima verosimilitud para la obtención de los factores. Señale de que forma varían los resultados.

BIBLIOGRAFÍA

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. y Tatham, R. L. (2006). Multivariate Data Analysis. Sixth Edition. Pearson, Prentice Hall. New Jersey.

Pérez López, C. (2005). Métodos estadísticos avanzados con SPSS. Thomson Paraninfo.

Pérez López, C. (2005). Técnicas estadísticas con SPSS12. Aplicaciones al análisis de datos. Pearson Alambra.

Pérez López, C. (2004). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS. Pearson Alhambra.

Pérez López, C. y Santín González, D. (2005). Data Mining. Soluciones con Enterprise Miner. RA-MA.

Pérez López, C. y Santín González, D. (2007). Minería de Datos. Técnicas y herramientas. Thomson Paraninfo.

CAPÍTULO XII
ANÁLISIS CLUSTER JERÁRQUICO

CÉSAR PÉREZ LÓPEZ
DANIEL SANTÍN GONZÁLEZ

12.1. PRINCIPIOS DEL ANÁLISIS CLUSTER

Hacer clasificaciones es una de las tareas más antiguas de la estadística. Es frecuente que cuando nos enfrentamos a un problema que afecta a las familias su forma de resolución, los mecanismos de atención o los medios necesarios para tener éxito en la intervención varíen según las circunstancias de cada una. Además, en la articulación de políticas de familia es necesario identificar aquellas «familias de riesgo» que en el futuro pueden ser beneficiarias de las ayudas. Dado que cada familia es diferente de otra es posible que *a priori* no conozcamos cuáles son los grupos de familias característicos de la población a la cual nos enfrentamos. El análisis cluster aborda la tarea de clasificar elementos (individuos, familias, receptores de ayudas, posibles beneficiarios, ayuntamientos, barrios, etc.) de acuerdo con sus características observadas. Podríamos resumir los *principios básicos del análisis cluster* (o de conglomerados) como sigue:

- El análisis cluster es un método estadístico multivariante de clasificación automática de datos.
- Su finalidad esencial es revelar concentraciones en los datos (casos o variables) para su agrupamiento eficiente en clusters (o conglomerados) según su homogeneidad.
- El agrupamiento puede realizarse tanto para casos como variables, pudiendo utilizarse variables cualitativas o cuantitativas.
- Los grupos de casos o variables se realizan basándose en la proximidad o lejanía de unos con otras, por lo tanto es esencial el uso adecuado del concepto de distancia.
- Es fundamental que los elementos dentro de un cluster sean homogéneos y lo más diferentes posibles de los contenidos en otros clusters.
- El análisis cluster es por tanto una técnica de clasificación, conociéndose también con el nombre de *taxonomía numérica*. Otros nombres asignados al mismo concepto son análisis de *conglomerados, análisis tipológico, clasificación automática* entre otros.

- El número de clusters no es conocido de antemano y los grupos se crean en función de la naturaleza de los datos.

Podíamos definir el análisis cluster como un método estadístico multivariante de clasificación automática que a partir de una tabla de datos (casos-variables), trata de situarlos en grupos homogéneos, conglomerados o clusters, no conocidos de antemano pero sugeridos por la propia esencia de los datos, de manera que los individuos que puedan ser considerados similares sean asignados a un mismo cluster, mientras que individuos diferentes (disimilares) se localicen en clusters distintos. La diferencia esencial con el análisis discriminante estriba en que en este último es necesario especificar previamente los grupos por un camino objetivo (técnica de clasificación *ad hoc*), ajeno a la medida de las variables en los casos de la muestra. El análisis cluster define grupos tan distintos como sea posible en función de los propios datos sin especificación previa de los citados grupos (técnica de clasificación *post hoc*).

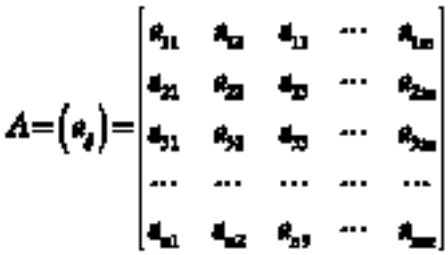
Para trabajar en análisis cluster es necesario tener presentes determinadas condiciones entre las que destacan las siguientes:

- Si las variables de aglomeración están en escalas muy diferentes será necesario estandarizar previamente las variables, o por lo menos trabajar con desviaciones respecto de la media.
- Es necesario observar también los valores atípicos y desaparecidos porque los métodos jerárquicos no tienen solución con valores perdidos y los valores atípicos deforman las distancias y producen clusters unitarios.
- También es nocivo para el análisis cluster la presencia de variables correlacionadas, de ahí la importancia del análisis previo de multicolinealidad.
- Si es necesario se realiza un análisis factorial previo y posteriormente se aglomeran las puntuaciones factoriales.
- La solución del análisis cluster no tiene porqué ser única, pero no deben encontrase soluciones contradictorias por distintos métodos.
- El número de observaciones en cada cluster debe ser relevante, ya que en caso contrario puede haber valores atípicos que difuminen la construcción de los clusters.
- Los conglomerados deben de tener sentido conceptual y no variar mucho al variar la muestra o el método de aglomeración.

- Los grupos finales serán tan distintos como permitan los datos. Con estos grupos se podrán realizar otros análisis: descriptivos para caracterizar los grupos y asignarles un nombre, discriminante, regresión logística, diferencias entre los grupos en función de la política de familia que se haya o esté aplicando, etc.

12.2. EL PROBLEMA MATEMÁTICO

De forma más general, podemos representar la tabla de datos (observaciones - variables) mediante la matriz siguiente:



Los individuos que forman parte del estudio, y que se intentan clasificar, vendrán caracterizados o definidos por diferentes valores obtenidos al medir determinadas variables sobre ellos, es decir, cada individuo poseerá un determinado valor para cada una de las variables que se traten en el estudio. De esta manera, si se consideran n individuos que se denotan por P_1, ..., P_n, y se consideran m variables, llamadas x_1, ..., x_m, los datos que definen a toda la muestra se pueden representar en la matriz de datos $A = (a_{ij})$, de modo que cada individuo aparece en cada una de las filas, y los valores que cada variable toma para cada individuo aparece en cada una de las columnas. Es decir, las puntuaciones que definen al individuo P_i serán los valores a_{i1}, a_{i2}, ..., a_{im}, Por tanto, los individuos se corresponden con las filas de la matriz y las variables con sus columnas.

Por otro lado, tendremos presente que un espacio métrico es un espacio en el que se ha definido una distancia (métrica o forma de medir). Si en un espacio métrico consideramos como sistema de ejes de coordenadas el definido por las variables objeto del estudio, se está en un espacio de tantas dimensiones como número de variables se considera, es decir, m dimensiones. Entonces, cada uno de los n individuos puede ser tomado como un punto en dicho espacio métrico dando lugar a una nube de n puntos. De este modo, cada uno de los valores a_{ij} (que representa la proporción de la variable x_j que entra a formar parte del individuo P_i), que definen a cada uno de los individuos se considerarán como las coordenadas del mismo.

Simétricamente, también puede considerarse un espacio métrico con el sistema de ejes coordenados definido por los *n* individuos y considerar cada una de las *m* variables como un punto de dicho espacio métrico dando lugar a una nube de *m* puntos. El objetivo del análisis cluster consiste en separar de alguna forma los puntos de estas nubes, de modo que se obtengan grupos de individuos o variables relativamente parecidos entre sí. Debido a este objetivo de separar los puntos es por lo que se recurre a un espacio métrico donde se tenga definida una forma de medir (métrica) a través de una *distancia* para comprender la separación.

Medidas de similitud

Según la clasificación de Sneath y Sokal existen cuatro grandes tipos de medidas de similitud.

- *Distancias*: se trata de las distintas medidas entre los puntos del espacio definido por los individuos. Se trata de las medidas inversas de las similitudes, es decir, disimilitudes. El ejemplo más clásico es la *distancia euclídea*.
- *Coeficientes de asociación*: se utilizan cuando trabajamos con datos cualitativos, aunque también se pueden aplicar a datos cuantitativos si se está dispuesto a sacrificar alguna información proporcionada por los individuos o las variables. Estas medidas son, básicamente, una forma de medir la concordancia o conformidad entre los estados de dos columnas de datos.
- *Coeficientes angulares*: se utilizan para medir la proporcionalidad e independencia entre los vectores que definen los individuos. El más común es el coeficiente de correlación aplicado a variables continuas.
- *Coeficientes de similitud probabilística*: miden la homogeneidad del sistema por particiones o subparticiones del conjunto de los individuos e incluyen información estadística. La idea de utilizar estos coeficientes se basa en relacionarlos con diferentes clasificaciones utilizando para ellas criterios de bondad o buenos ajustes estadísticos. Las principales propiedades de estos coeficientes es que son aditivos, se distribuyen como la *Chi cuadrado* y son probabilísticas. Esta última propiedad permite, en aquellos casos en que es posible, establecer una hipótesis nula y contrastarla por los métodos estadísticos tradicionales.

Dado que no es el objetivo de este capítulo realizar una revisión extensiva de las numerosas definiciones de medidas a continuación sólo se

presentan los ejemplos más característicos de cada uno de estos tipos de medidas de similitud.

$$\text{Distancias}\begin{cases}\text{Distancia euclídea al cuadrado } d(i,j)^2=\sum_k\left(x_{ik}-x_{jk}\right)^2\\ \text{Distancia euclídea } d(i,j)=\sqrt{\sum_k\left(x_{ik}-x_{jk}\right)^2}\\ \text{Distancia de Minkowski } d_q(i,j)=\left(\sum_k\left|x_{ik}-x_{jk}\right|^q\right)^{1/q}\\ \text{Distancia City-Block o de Manhatan } d_1(i,j)=\sum_k\left|x_{ik}-x_{jk}\right|\\ \text{Distancia de Chebychev } d_\infty(i,j)=Max_k\left(\left|x_{ik}-x_{jk}\right|\right)\\ \text{Distancia de Canberra } d_{Can}(i,j)=\sum_k\frac{\left|x_{ik}-x_{jk}\right|}{\left(x_{ik}+x_{jk}\right)}\end{cases}$$

Se observa que la distancia euclídea al cuadrado entre dos individuos se define como la suma de los cuadrados de las diferencias de todas las coordenadas de los dos puntos. La distancia euclídea se define como la raíz cuadrada positiva de la distancia anterior. La distancia de Minkowski es una distancia genérica que da lugar a otras distancias en casos particulares y se define como la raíz q-ésima de la suma de las potencias q-ésimas de las diferencias, en valor absoluto, de las coordenadas de los dos puntos considerados. La distancia City-Block o distancia de Manhatan, es un caso particular de la distancia o medida de Minkowski cuando $q = 1$ y resulta ser la suma de las diferencias, en valor absoluto, de todas las coordenadas de los dos individuos cuya distancia se calcula. El valor de esta medida es cero para la similitud perfecta y aumenta a medida que los objetos son más disimilares. La distancia de Chebychev se define como el caso límite de la medida de Minkowski para q tendiendo a infinito, es decir, es el máximo de las diferencias absolutas de los valores de todas las coordenadas. La distancia Canberra es una modificación de la distancia City-Block que es sensible a proporciones y no sólo a valores absolutos.

Los coeficientes de asociación suelen utilizarse para el caso de variables cualitativas, y en general para el caso de datos binarios (o dicotómicos), que son aquéllos que sólo pueden presentar dos opciones (blanco - negro, sí - no, hombre - mujer, verdadero - falso, etc.). En este caso existen diferentes medidas de proximidad o similitud, que se verán a continuación, partiendo de una tabla de frecuencias 2x2 en la que se representa el número de elementos

de la población en los que se constata la presencia o ausencia del carácter (variable cualitativa) en estudio.

Variable 1 → Variable 2 ↓	Presencia	Ausencia
Presencia	a	b
Ausencia	c	d

Coeficientes de asociación

Jaccard-Sneath $S_J = \frac{a}{(a+u)} = \frac{a}{(a+b+c)}$

Coeficiente de emparejamiento simple

$S_{SM} = \frac{m}{(m+u)} = \frac{m}{n} = \frac{(a+d)}{(a+b+c+d)}$

Coeficiente de Yule $S_Y = \frac{(ad-bc)}{(ad+bc)}$

El *coeficiente de Jaccard - Sneath* es uno de los coeficientes más sencillos, que no tiene en cuenta los emparejamientos negativos, y se define como el número de emparejamientos positivos entre la suma de los emparejamientos positivos y los desacuerdos. A partir de su expresión se deduce que S_J tiende a cero cuando *a/u* tiende a cero, esto es, S_J es cero cuando el número de emparejamientos positivos coincide con el de desacuerdos. también S_J tiende a uno cuando *u* tiende a cero, es decir, S_J vale uno cuando no hay desacuerdos. El coeficiente de Yule varía entre +1 y -1. El *coeficiente de emparejamiento simple* se define como el cociente entre el número de emparejamientos y el número total de casos considerados. De su expresión se deduce:

$$S_{SM} \to 0 \text{ si } \frac{m}{u} \to 0 \text{ y } S_J \to 1 \text{ si } \frac{a}{m} \to 1.$$

En el caso de los *coeficientes angulares* su campo de variación está entre -1 y +1. Los valores cercanos a 0 indican disimilitud entre los individuos y los valores que se acercan a +1 o a -1 indican similitud positiva o negativa respectivamente. El cálculo de este coeficiente entre los individuos *i* y *j* se realiza en función de X_i y X_j que son las medias correspondientes a los individuos *i* y *j*.

Coeficientes angulares

Coeficiente de correlación $r_{ij} = \dfrac{\sum_k (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\left[\sum_k (x_{ik} - \bar{x}_i)^2 \sum_k (x_{jk} - \bar{x}_j)^2\right]^{1/2}}$

Distancia del coseno $\cos\alpha_{ij} = \dfrac{\left(\sum_k x_{ik} x_{jk}\right)}{\left[\sum_k (x_{ik})^2 \sum_k (x_{jk})^2\right]^{1/2}}$

Los *coeficientes de similitud probabilística* calculan la probabilidad acumulada de que un par de individuos *i* y *j*, sean tan similares, o más, que lo que empíricamente se puede afirmar sobre la base de la distribución observada.

Para el caso de variables cualitativas y en general para el caso de datos binarios o dicotómicos existen varias medidas de similaridad adicionales que se muestran en la tabla siguiente:

Russel y Rao	$RR_{ij} = \dfrac{a}{a+b+c}$	Sokal y Sneath	$SS_{ij} = \dfrac{2(a+d)}{2(a+d)+b+c}$
Pareos simples	$PS_{ij} = \dfrac{a+d}{a+b+c+d}$	Rogers y Tanimoto	$RT_{ij} = \dfrac{a+d}{a+d+2(b+c)}$
Jaccard	$J_{ij} = \dfrac{a}{a+b+c}$	Sokal y Sneath(2)	$SS2_{ij} = \dfrac{a}{a+2(b+c)}$
Dice y Sorensen	$D_{ij} = \dfrac{2a}{2a+b+c}$	Kulczynski	$K_{ij} = \dfrac{a}{b+c}$

Hay otro grupo de medidas denominadas medidas de similaridad para probabilidades condicionales, entre las que destacan las siguientes:

Kulczynski (medida 2)	$K2_{ij} = \dfrac{a/(a+b)+a/(a+c)}{2}$
Sokal y Sneath (medida 4)	$SS4_{ij} = \dfrac{a/(a+b)+a/(a+c)+d/(b+d)+d/(c+d)}{4}$
Hamann	$H_{ij} = \dfrac{(a+d)-(b+c)}{a+b+c+d}$

También suele considerarse un subgrupo de medidas denominadas de predición entre las que se encuentran la D_{xy} de Anderberg, la Y_{xy} de Yule y la Q_{xy} de Yule, que se definen como sigue:

$$D_w = \frac{\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)-\max(a+c,b+d)-\max(a+b,c+d)}{2(a+b+c+d)}$$

$$Y_w = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \qquad Q_w = \frac{ad-bc}{ad+bc}$$

Por último, se usan otras medidas binarias, entre las que destacan las siguientes:

Ochiai	$O_w = \sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}}$	*Sokal y Sneath (5)*	$SS_w = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Sokal y Sneath (3)	$SS_w = \frac{a+d}{b+c}$	*Correlación phi*	$\phi_w = \frac{ad-bc}{(a+b)(a+c)(b+c)(c+d)}$
Euclides binario	$EB_w = \sqrt{b+c}$	*Diferencia de forma*	$DF_w = \frac{(a+b+c+d)(b+c)-(b-c)^2}{(a+b+c+d)^2}$
Euclides binario²	$EB_w^2 = b+c$	*Varianza diferencial*	$V_w = \frac{b+c}{4(a+b+c+d)}$
Dispersión	$D_w = \frac{ad-bc}{(a+b+c+d)^2}$	*Diferencia de tamaño*	$T_w = \frac{(b-c)^2}{(a+b+c+d)^2}$
Lance y Williams	$LW_w = \frac{b+c}{2a+b+c}$	*Diferencia de patrón*	$P_w = \frac{bc}{(a+b+c+d)^2}$

12.3. TÉCNICAS EN EL ANÁLISIS CLUSTER

Ya sabemos que el análisis de conglomerados o análisis cluster es un conjunto de métodos y técnicas estadísticas que permiten describir y reconocer diferentes agrupaciones que subyacen en un conjunto de datos, es decir, permiten clasificar, o dividir en grupos más o menos homogéneos, un conjunto de individuos que están definidos por diferentes variables. El objetivo principal del análisis de conglomerados consiste, por tanto, en conseguir una o más particiones de un conjunto de individuos en base a determinadas características de los mismos. Estas características estarán definidas por las puntuaciones que cada uno de ellos tiene con relación a diferentes variables. Así, se podrá decir que dos individuos o familias son similares si pertenecen a la misma clase, grupo, conglomerado o cluster. Si se consigue este objetivo, se tendrá que todos los individuos que están contenidos en el mismo conglomerado se parecerán entre sí, y serán diferentes de los individuos que pertenecen a otro conglomerado. Por tanto, los miembros de un conglomerado gozarán de unas características comunes que los diferencian de los miembros de otros conglomerados. Estas características deberán, por la definición del objetivo a conseguir, ser genéricas, y es claro que difícilmente una única característica podrá definir un conglomerado.

El método para ejecutar un análisis de conglomerados comienza con la selección de los individuos objeto del estudio, incluyendo en algunas casos su codificación a partir de las variables o caracteres que los definen y su transformación adecuada para someterlos al análisis si es necesario (tipificación de variables, desviaciones respecto de la media, etc.). A continuación se determina la matriz de disimilitudes definiendo las distancias, similitudes o disimilitudes de los individuos vistas en el apartado anterior. Una vez determinadas las disimilitudes de los individuos, se procede a ejecutar el algoritmo que formará las diferentes agrupaciones o conglomerados de individuos. Determinada ya la clasificación, el paso siguiente consiste en obtener una representación gráfica de los conglomerados obtenidos, de modo que se puedan visualizar los resultados alcanzados. Este proceso se lleva a cabo mediante un dendrograma. Conseguido el propósito de la clasificación, la última fase a llevar a cabo es la interpretación de los resultados obtenidos.

Los diferentes métodos de análisis de conglomerados surgen de las diferentes formas de llevar a cabo la agrupación de los individuos, es decir, dependiendo del algoritmo que se utilice para llevar a cabo la agrupación de individuos o familias, se obtienen diferentes métodos de análisis de conglomerados. Una clasificación de los métodos de análisis de conglomerados basada en los algoritmos de agrupación de individuos podría ser la siguiente:

- *Métodos Aglomerativos-Divisivos*: un método es aglomerativo si considera tantos grupos como individuos y sucesivamente va fusionando los dos grupos más similares, hasta llegar a una clasificación determinada; mientras que un método es divisivo si parte de un solo grupo formado por todos los individuos, de modo que en cada etapa va separando individuos de los grupos establecidos anteriormente, formándose así nuevos grupos.

- *Métodos Jerárquicos-No jerárquicos*: un método es jerárquico si consiste en una secuencia de g+1 clusters: $G_0, \ldots, G_g$ en la que G_0 es la partición disjunta de todos los individuos y G_g es el conjunto partición. El número de partes de cada una de las particiones disminuye progresivamente, lo que hace que éstas sean cada vez más amplias y menos homogéneas. Por el contrario, un método se dice no jerárquico cuando se forman grupos homogéneos sin establecer relaciones de orden o jerárquicas entre dichos grupos.

- *Métodos Solapados-Exclusivos*: un método es solapado si admite que un individuo pueda pertenecer a dos grupos simultáneamente en alguna de las etapas de clasificación, mientras que se dice exclusivo si ningún individuo puede pertenecer simultáneamente a dos grupos en la misma etapa.

- Método *Secuenciales-Simultáneos*: un método es secuencial si a cada grupo se le aplica el mismo algoritmo en forma recursiva, mientras que los métodos simultáneos son aquellos en los que la clasificación se logra por una simple y no reiterada operación sobre los individuos.

- *Métodos Monotéticos-Politéticos*: un método se dice monotético si está basado en una característica única de los objetos a clasificar; mientras que es politético si se basa en varias características de los mismos, sin exigir que todos los objetos las posean, aunque sí las suficientes como para poder justificar la analogía entre los miembros de una misma clase.

- *Métodos Directos-Iterativos*: un método es directo si utiliza algoritmos en los que una vez asignado un individuo a un grupo ya no se saca del mismo, mientras que los métodos iterativos corrigen las asignaciones previas volviendo a comprobar en posteriores iteraciones si la asignación de un individuo a un conglomerado es óptima, llevando a cabo un nuevo reagrupamiento de los individuos si es necesario.

- *Métodos Ponderados-No ponderados*: los métodos no ponderados son aquellos que establecen el mismo peso a todas las características de los individuos a clasificar; mientras que los ponderados hacen recaer mayor peso en determinadas características.

- *Métodos Adaptativos-No adaptativos*: Los métodos no adaptativos son aquellos para los que el algoritmo utilizado se dirige hacia una solución en la que el método de formación de conglomerados es fijo y está predeterminado, mientras que los adaptativos (menos utilizados) son aquellos que de alguna manera aprenden durante el proceso de formación de los grupos y modifican el criterio de optimización o la medida de similitud a utilizar.

12.4. CONGLOMERADOS JERÁRQUICOS, SECUENCIALES, AGLOMERATIVOS Y EXCLUSIVOS (S.A.H.N.)

Los *métodos de análisis de conglomerados* que más se usan son los que son a la vez secuenciales, aglomerativos, jerárquicos y exclusivos, y que reciben el acrónimo, en lengua inglesa, de S.A.H.N. (*Sequential, Agglomerative, Hierarchic* y *Nonoverlaping*). En todos los *métodos de tipo S.A.H.N.* se siguen dos pasos fundamentales en el proceso de elaboración de los conglomerados. El primero de ellos es que los coeficientes de similitud o disimilitud entre los nuevos conglomerados establecidos y los candidatos potenciales a ser

admitidos se recalcula en cada etapa, y el otro es el criterio de admisión de nuevos miembros a un conglomerado ya establecido. En los párrafos siguientes se estudian los diferentes métodos de análisis de conglomerados de tipo S.A.H.N. a cuya implementación podemos acceder en Excel mediante XLSTAT → Clasificación Ascendente Jerárquica (CAJ) (Figura 12-1).

FIGURA 12-1

Método de unión simple (Single Linkage Clustering), entorno o vecino más cercano (Nearest Neighbour) o método del mínimo (Minimum Method)

Este método relaciona un elemento con un grupo si tiene la mayor similitud con cualquiera de los elementos individuales de ese grupo. Este tipo de unión permite que se pueda realizar con sólo inspeccionar la matriz de similitudes. Los dos primeros casos que se combinan son aquellos cuya distancia es la menor o cuya similitud es máxima. La distancia entre el nuevo conglomerado y un caso individual se calcula como la mínima distancia entre el caso individual y un caso del conglomerado. La distancia entre dos casos que no han sido unidos no cambia. En cada caso, la dis-

tancia entre dos conglomerados se toma como la distancia entre dos puntos más cercanos. Este método utiliza la distancia:

$$d(h_k, h_i \cup h_j) = \text{Min}\left(d(h_i, h_k), d(h_j, h_k)\right)$$

Método de la distancia máxima o método del máximo (Complete Linkage Clustering, Furthest Neighbour o Maximum Method)

En este método la similitud de un elemento con un grupo se calcula como la similitud de dicho elemento con el individuo más alejado de ese grupo. La distancia entre dos clusters se calcula como la distancia entre sus dos puntos más alejados. Este método se define mediante la distancia siguiente:

$$d(h_k, h_i \cup h_j) = \text{Max}\left(d(h_i, h_k), d(h_j, h_k)\right)$$

Método de la media o de la distancia promedio no ponderado (Weighted Pair Groups Method Using Arithmetic Averages WPGMW)

Este método pondera los nuevos miembros admitidos en un conglomerado con el mismo peso que los existentes hasta entonces. El método combina conglomerados de modo que la distancia media entre todos los casos en el conglomerado resultante sea la menor posible. Así, la distancia entre dos conglomerados se toma como la media de las distancias entre todos los posibles pares de casos en el conglomerado resultante. Este método usa la distancia:

$$d(h_k, h_i \cup h_j) = \left(\frac{1}{2}\right) d(h_i, h_k) + \left(\frac{1}{2}\right) d(h_j, h_k)$$

Método de la media ponderada o de la distancia Promedio Ponderado (Group Average o Unweighted Pair Groups Method Using Arithmetic Averages UPGMA)

En este método, similar al de la media, la distancia entre dos conglomerados se define como la media de las distancias entre todos los pares de casos en los que un miembro del par es de cada uno de los conglomerados. La distancia se define ponderando respecto a n_i y n_j; es decir, ponderando con respecto al número de individuos de h_i y *de* h_j de la siguiente forma:

$$d(h_k, h_i \cup h_j) = \left(\frac{n_i}{(n_i + n_j)}\right) d(h_i, h_k) + \left(\frac{n_j}{(n_i + n_j)}\right) d(h_j, h_k)$$

Método de la mediana o de la distancia mediana (Weighted Pair Group Centroid Method WPGMC)

En este método los dos conglomerados que están siendo combinados pesan lo mismo en el cálculo del centroide y es indiferente el número de casos de cada uno. Esto permite que conglomerados pequeños tengan igual efecto en la caracterización que los conglomerados grandes con los que están siendo mezclados. Este método utiliza sólo la distancia euclídea al cuadrado definiéndose su distancia como sigue.

$$d(h_i, h_j \cup h_j) = \left(\frac{1}{2}\right) d(h_i, h_k) + \left(\frac{1}{2}\right) d(h_j, h_k) - \left(\frac{1}{4}\right) d(h_i, h_j)$$

Método del Centroide o de la Distancia Prototipo (Unweighted Pair Group Centroid Method UPGMC)

Este método calcula la distancia entre dos conglomerados como la distancia entre sus medias para todas las variables.

Una desventaja del método es que la distancia con la que los conglomerados se combinan disminuye de un paso al siguiente. Es una propiedad no deseable pues los conglomerados mezclados en etapas posteriores son menos similares que los mezclados en etapas anteriores. El centroide de un conglomerado mezclado es una combinación ponderada de los centroides de los dos conglomerados individuales, donde los pesos son proporcionales a los tamaños de los conglomerados. Este método es similar al anterior, pero en él se hace intervenir el número de individuos de h_i y de h_j, que son n_i y n_j respectivamente. La distancia que se define es la siguiente:

$$d(h_i, h_j \cup h_j) = \left(\frac{n_i}{(n_i + n_j)}\right) d(h_i, h_k) + \left(\frac{n_j}{(n_i + n_j)}\right) d(h_j, h_k) - \left(\frac{n_i n_j}{(n_i + n_j)^2}\right) d(h_i, h_j)$$

Método de Ward o de mínima varianza

Para este método se considera la distancia euclídea al cuadrado como medida de disimilitud. Llamando $d(x_i, x_j) = \|x_i - x_j\|^2$ a la distancia entre los puntos x_i y x_j, la varianza total (o inercia) del conjunto de puntos es la cantidad dada por la expresión $I = \sum_i m_i \|x_i - G\|^2$ siendo G el centro de gravedad de los puntos dados, con masas respectiva m_i. Si existe una partición del con-

junto de individuos en *q* conglomerados, el q-ésimo conglomerado tiene como centro de gravedad a G_q y masa m_q.

Entonces la inercia se puede descomponer como la suma de la varianza que existe dentro de los conglomerados y la que hay entre unos conglomerados y otros, de la forma $I = \sum_q m_q \|G_q - G\|^2 + \sum_q \sum_{i \in q} m_i \|x_i - G_q\|^2$. Si x_i y x_j son dos elementos de masas m_i y m_j respectivamente, que se unen en un elemento x de masa $m = m_i + m_j$, con $x = (m_i x_i + m_j x_j) / (m_i + m)$, podemos descomponer la varianza I_{ij} de x_i y x_j con respecto a G por la ecuación $I_{ij} = m_i \|x_i - x\|^2 + m_j \|x_j - x\|^2 + m\|x - G\|^2$. El último término es el único que permanece constante si se cambian x_i y x_j por su centro de gravedad x. La reducción en la varianza es $\Delta I_{ij} = m_i \|x_i - x\|^2 + m_j \|x_j - x\|^2$. Reemplazando x por su valor como función de x_i y x_j, tenemos:

$$\Delta I_{ij} = \left(\frac{(m_i m_j)}{(m_i + m_j)}\right) \|x_i - x\|^2 = \left(\frac{(m_i m_j)}{(m_i + m_j)}\right) d(x_i - x_j)^2$$

El método que se sigue para hacer conglomerados con este método consiste en encontrar los individuos x_i y x_j con la condición de que hagan mínima ΔI_{ij}, en lugar de ser los individuos más cercanos. Por tanto, puede considerarse a ΔI_{ij} como un nuevo índice de disimilitud.

Por medio de este método, los individuos con menor peso son los que más pronto se unen. El cuadrado de la distancia de un punto z a un centro de conglomerados x, se puede escribir en función de las distancias a los puntos x_i y x_j:

$$d(x - z)^2 = \left(\frac{1}{(m_i + m_j)\left(m_i d(x_i, z)^2\right)}\right) + m_j d(x_j, z)^2 - \left(\frac{(m_i m_j)}{(m_i + m_j)}\right) d(x_i, x_j)^2$$

Fórmula de Lance y Williams para la distancia entre grupos

Matemáticamente, Lance y Williams desarrollaron una fórmula general que puede ser utilizada para describir los distintos tipos de enlaces de los métodos jerárquicos aglomerativos. La *fórmula de Lance y Williams para la distancia entre grupos* es la siguiente:

$$D_{k(i,j)} = a_i D_{ki} + a_j D_{kj} + b D_{ij} + g |D_{ki} - D_{kj}|$$

donde D_{ij} es la distancia entre los grupos *i* y *j*, y a, b y g son los tres parámetros del modelo. Se observa lo siguiente:

$a_i = a_j = 1/2$, $b = 0$ y $g = -1/2 \Rightarrow$ enlace simple

$a_i = a_j = 1/2$, $b = 0$ y $g = 1/2 \Rightarrow$ enlace completo

$a_i = a_j = 1/2$, $b = -1/4$ y $g = 0 \Rightarrow$ método de la mediana

$\alpha_i = \frac{n_i}{n_i + n_j}$, $\alpha_j = \frac{n_j}{n_i + n_j}$, $b = -a_i a_j$ y $g = 0 \Rightarrow$ enlace centroide

$\alpha_i = \frac{n_i}{n_i + n_j}$, $\alpha_j = \frac{n_j}{n_i + n_j}$, $b = g = 0 \Rightarrow$ enlace promedio

$\alpha_i = \frac{n_k + n_i}{n_k + n_i + n_j}$, $\alpha_j = \frac{n_k + n_j}{n_k + n_i + n_j}$, $\beta = \frac{-n_k}{n_k + n_i + n_j}$ y $\gamma = 0 \Rightarrow$

Ward

$a_i + a_j + b = 1$, $a_i = a_j$, $b < 1$ y $g = 0$ ⌋ método flexible (cuádruple restricción)

El último método (*cuádruple restricción*) consiste en utilizar la forma de Lance y Williams variando los coeficientes según las necesidades del clasificador, pero respetando las cuatro restricciones impuestas.

Los métodos de clusters jerárquicos, por la laboriosidad de los cálculos, no resultan prácticos para procesar grandes ficheros de datos. En estos casos, puede ser aconsejable realizar un análisis previo no jerárquico, que proporcione un número preliminar razonable de clusters (en lugar de individuos) que servirán luego de partida para su posterior clasificación jerárquica.

Como resumen, los métodos jerárquicos producen resultados más ricos que los no jerárquicos. Con un solo análisis se obtiene una configuración de grupos en cada nivel de clasificación. Los mismos indicadores que en clasificación no jerárquica valoraban la adecuación del número de clusters (Criterio cúbico de clusters, Pseudo F, etc.) permiten detectar aquí el nivel jerárquico en que la separación de los grupos formados es más ostensible.

12.5. DENDOGRAMA

Es habitual en la investigación la necesidad de clasificar los datos en grupos con estructura arborescente de dependencia, de acuerdo con diferentes niveles de jerarquía. Partiendo de tantos grupos iniciales como individuos se estudian, se trata de conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clusters los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntarán hasta llegar al cluster final que contiene todos los casos analizados.

La representación gráfica de estas etapas de formación de grupos, a modo de árbol invertido, se denomina *dendograma* y se representa a continuación (Figura 12-2):

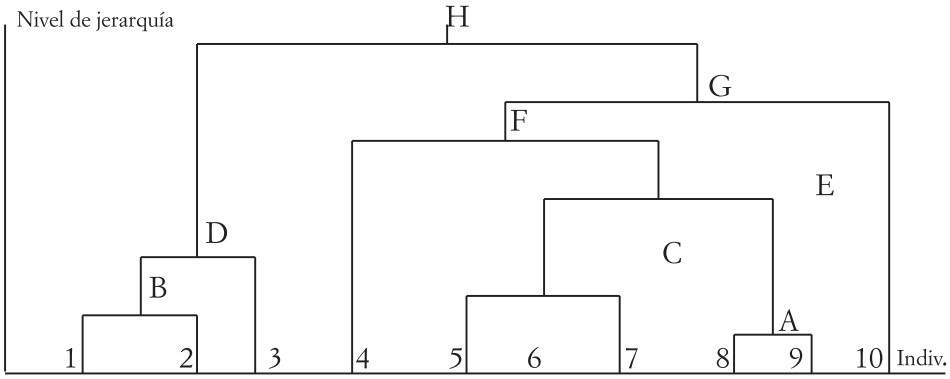


FIGURA 12-2

La figura, que corresponde a un estudio de los individuos, muestra cómo el 8 y el 9 se agrupan en un primer cluster (A). En un nivel inmediatamente superior, se unen los individuos 1 y 2 (cluster B); y enseguida los 5, 6, y 7 (C). Un paso siguiente engloba el cluster B con el individuo 3 (D); y así sucesivamente hasta que todos ellos quedan estructurados al conseguir, en el nivel más alto, el cluster total (H) que reúne los 10 casos.

EJERCICIO 12-1. El fichero REGIONES contiene información socio-educativa referente a 19 regiones. Se pretende indagar acerca de cuantos grupos de regiones existen para posteriormente diseñar un sistema de ayudas

a) **Realizar un análisis cluster jerárquico determinado por la distancia de Chebychev y el método de Ward.**

b) **Asumiendo que se formaran tres grupos, realizar las estadísticas descriptivas de éstos.**

c) **Representar el dendograma de la clasificación.**

Para realizar este cálculo acudiremos a XLSTAT → Clasificación Ascendente Jerárquica (CAJ) y seleccionaremos las opciones que se muestran en la figura 12.3.

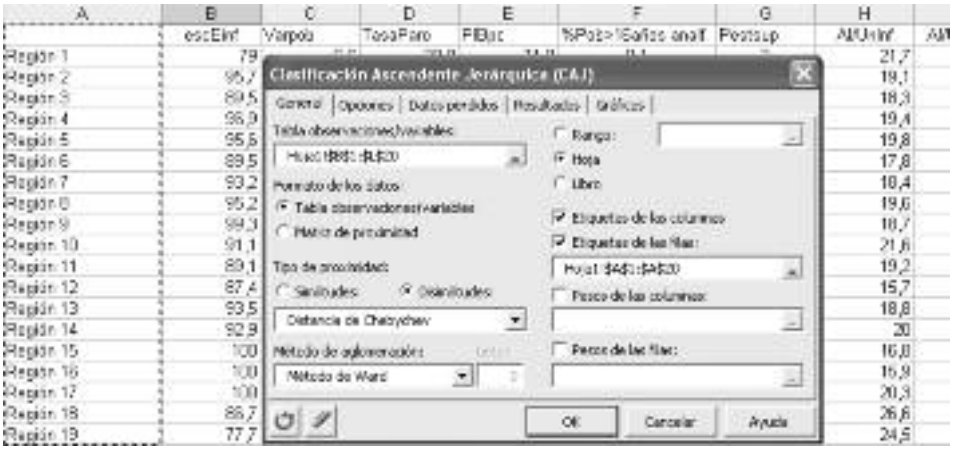


FIGURA 12-3

En el análisis cluster jerárquico no es necesario especificar cuantos grupos se formarán. Sin embargo podemos, una vez definidos los grupos que a nuestro juicio caracterizan las observaciones, calcular sus estadísticas descriptivas. Para ello la pestaña opciones (Figura 12-4) permite hacer el análisis agrupando por filas (como en nuestro caso y como en la mayoría de análisis en los que se utiliza el cluster) y establecer un truncamiento o número de grupos que se forman a partir de los datos. Este número se puede definir de forma automática, especificando el número de clases o estableciendo un nivel de altura en el proceso de unión de los casos.

FIGURA 12-4

Para solicitar que el análisis elabore el dendograma debemos acceder a la pestaña *Gráficos* (figura 12-5). En ella podemos establecer fundamentalmente si queremos una disposición horizontal o vertical del gráfico.

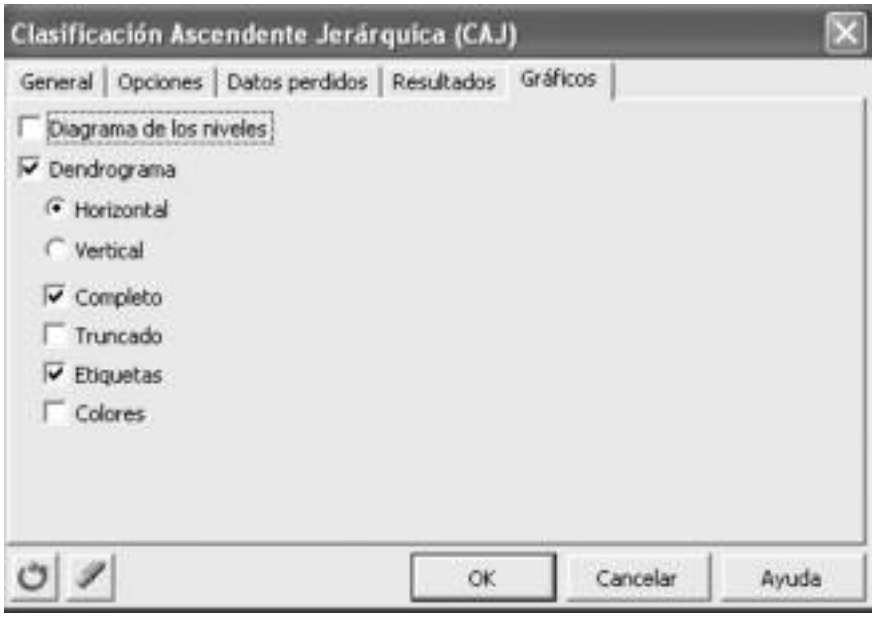


FIGURA 12-5

Una vez realizado el análisis podemos visualizar el dendograma (Figura 12-6), los grupos formados que definen los individuos (Figura 12-7) y las estadísticas descriptivas de los tres grupos formados (Figura 12.8) además de muchos otros resultados.

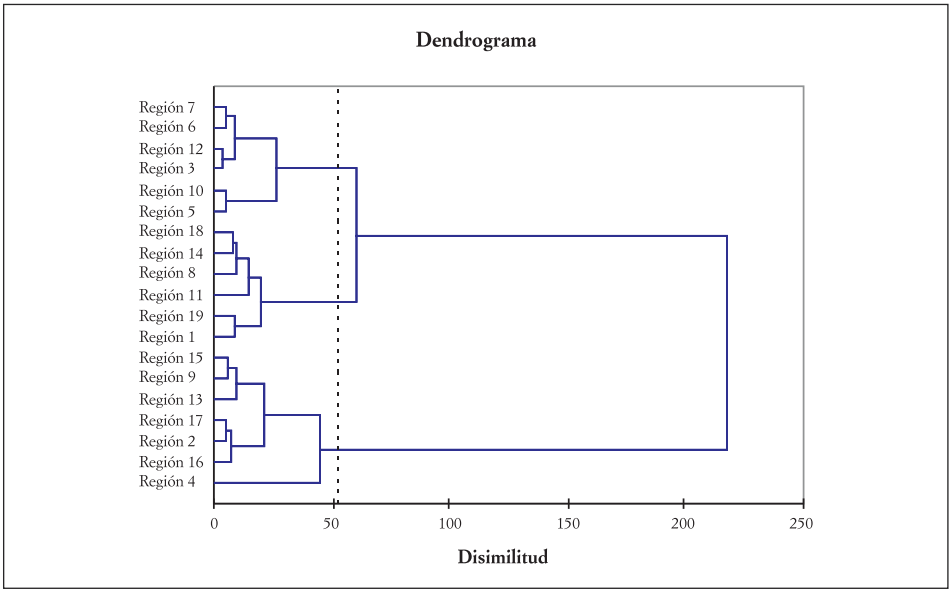


FIGURA 12-6

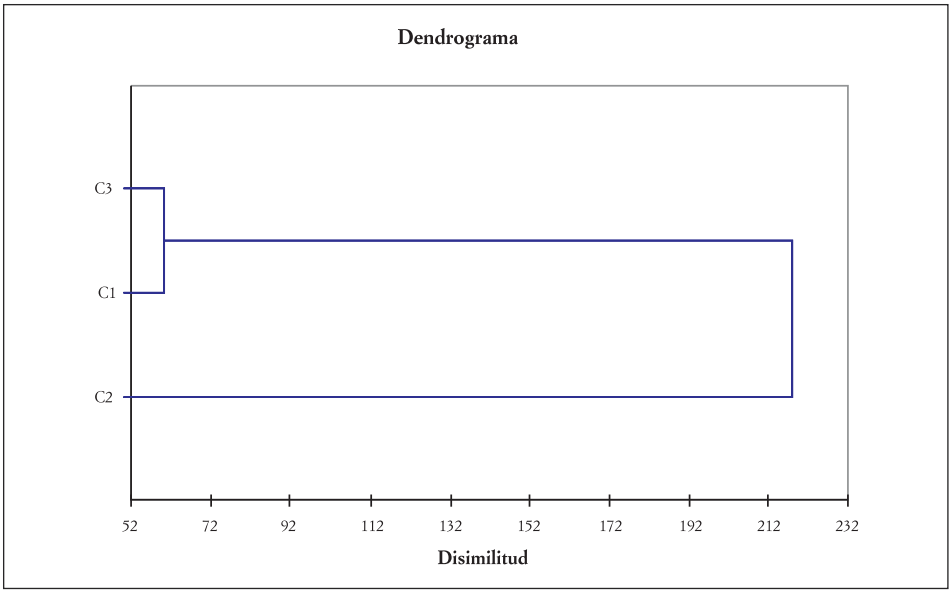


FIGURA 12-7

Cluster	1	2	3
escEinf	86,767	97,914	91,050
Varpob	7,267	4,100	1,533
TasaParo	23,767	13,300	18,733
PIBpc	75,017	121,957	92,717
%Pob>16años analf.	8,567	,943	2,867
Pestsup	2,717	5,257	3,417
Al/UnInf	21,933	18,429	18,600
Al/UnPrim	22,333	18,914	17,867
Al/UnEsp	6,983	5,414	5,317
%Pobjoven	20,805	16,850	17,607

FIGURA 12-8

EJERCICIO 12-2. Repita el ejercicio anterior utilizando la Distancia euclídea y el tipo de enlace medio. ¿Cree que son consistentes los resultados?

BIBLIOGRAFÍA

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. y Tatham, R. L. (2006). Multivariate Data Analysis. Sixth Edition. Pearson, Prentice Hall. New Jersey.

Pérez López, C. (2005). Métodos estadísticos avanzados con SPSS. Thomson Paraninfo.

Pérez López, C. (2005). Técnicas estadísticas con SPSS12. Aplicaciones al análisis de datos. Pearson Alambra.

Pérez López, C. (2004). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS. Pearson Alhambra.

Pérez López, C. y Santín González, D. (2005). Data Mining. Soluciones con Enterprise Miner. RA-MA.

Pérez López, C. y Santín González, D. (2007). Minería de Datos. Técnicas y herramientas. Thomson Paraninfo.

CAPÍTULO XIII
CORRESPONDENCIAS SIMPLES Y MÚLTIPLES

CÉSAR PÉREZ LÓPEZ
DANIEL SANTÍN GONZÁLEZ

13.1. ANÁLISIS DE CORRESPONDENCIAS

Hemos visto en un capítulo anterior que el análisis factorial, al igual que el análisis en componentes principales, son técnicas multivariantes que persiguen *reducir la dimensión de una tabla de datos* formada por *variables cuantitativas*. Si las variables fuesen *variables cualitativas*, estaríamos ante un análisis de correspondencias.

Cuando se estudia conjuntamente el comportamiento de dos variables cualitativas estamos ante el *análisis de correspondencias simples*, pero este análisis puede ser generalizado para el caso en que se dispone de un número de variables cualitativas mayor que dos, en cuyo caso estamos ante el *análisis de correspondencias múltiples*. En el caso de correspondencias simples los datos de las dos variables cualitativas pueden representarse en una tabla de doble entrada, denominada *tabla de contingencia*. En el caso de las correspondencias múltiples la tabla de contingencia de doble entrada pasa a ser una hipertabla en tres o más dimensiones, difícil de representar y que suele sintetizarse en la denominada *tabla disyuntiva completa* o *tabla de Burt*.

El objetivo del análisis de correspondencias es establecer relaciones entre variables no métricas enriqueciendo la información que ofrecen las tablas de contingencia, que sólo comprueban si existe alguna relación entre las variables (test de la *Chi-cuadrado*, etc.) y la intensidad de dicha relación (test *V* de Cramer, etc.).

El análisis de correspondencias revela además en qué grado contribuyen a esa relación detectada los distintos valores de las variables, información que suele ser proporcionada en modo gráfico (valores asociados próximos).

Podríamos sintetizar diciendo que el análisis de correspondencias busca como objetivo el estudio de la asociación entre las categorías de múltiples variables no métricas, pudiendo obtenerse un mapa perceptual que ponga de manifiesto esta asociación en modo gráfico.

13.2. ANÁLISIS DE CORRESPONDENCIAS SIMPLES ACS

El *análisis factorial de correspondencias simples* está particularmente adaptado para tratar tablas de contingencia, representando los efectivos existentes en las múltiples modalidades (categorías) combinadas de dos caracteres (variables cualitativas).

En el análisis de correspondencias simples se parte de una tabla de contingencia en la que se cruza el carácter I con modalidades desde $i = 1$ hasta $i = n$ (en filas), con el carácter J con modalidades desde $j = 1$ hasta $j = p$ (en columnas). Se representa el número de unidades estadísticas que pertenecen simultáneamente a la modalidad i del carácter I y a la modalidad j del carácter J mediante k_{ij}. En este caso, la distinción entre observaciones y variables en el cuadro de doble entrada es artificial, pero, por similitud con componentes principales, suele hablarse a veces de individuos u observaciones cuando nos refiramos al conjunto de las modalidades del carácter I (filas), y de variables cuando nos referimos al conjunto de las modalidades del carácter J (columnas).

La tabla de datos (k_{ij}) es una matriz K de orden (n, p) donde k_{ij} representa la frecuencia absoluta de asociaciones entre los elementos i y j, es decir el número de veces que se presentan simultáneamente las modalidades i y j de los caracteres I y J. El método buscado para el análisis factorial de correspondencias simple deberá ser simétrico con relación a las líneas y columnas de la tabla de contingencia (para estudiar las relaciones en el interior de los conjuntos I y J) y deberá permitir comparar las distribuciones de frecuencias de las dos características (para estudiar las relaciones entre los conjuntos I y J).

En cuanto a los objetivos generales del análisis de correspondencias simple, esencialmente se trata de estudiar las relaciones existentes en el interior del conjunto de modalidades del carácter I y las relaciones existentes en el interior del conjunto de modalidades del carácter J. Simultáneamente, también hay que estudiar las relaciones existentes entre las modalidades del carácter I y las modalidades del carácter J.

Para comparar dos líneas entre sí (filas o columnas) en una tabla de contingencia, no interesan los valores brutos sino los porcentajes o distribuciones condicionadas. En una tabla de contingencia, el análisis buscado debe trabajar no con los valores brutos k_{ij} sino con *perfiles* o porcentajes. No interesa poner de manifiesto las diferencias absolutas que existen entre dos líneas, sino que los elementos i,i' (j,j') se consideran semejantes si presentan la misma distribución condicionada.

Una primera caracterización de las modalidades i del carácter I (variables i) puede hacerse a partir del peso relativo (expresado en tanto por

uno) de cada modalidad del carácter *J* en la modalidad *i* denominado *perfil de la variable i* o distribución de frecuencias condicionada del carácter *J* para *I = i*:

$$\frac{k_{i1}}{k_{i.}}, \frac{k_{i2}}{k_{i.}}, \ldots, \frac{k_{ip}}{k_{i.}} \qquad k_{i.} = \sum_{j=1}^{p} k_{ij} = \text{efectivo total de la fila } i.$$

De modo análogo la caracterización de las modalidades *j* del carácter *J* (observaciones *j*) puede hacerse a partir del peso relativo (expresado en tanto por uno) de cada modalidad del carácter *I* en la modalidad *j* denominado *perfil de la observación j* o distribución de frecuencias condicionada del carácter *I* para *J = j*:

$$\frac{k_{1j}}{k_{.j}}, \frac{k_{2j}}{k_{.j}}, \ldots, \frac{k_{nj}}{k_{.j}} \qquad k_{.j} = \sum_{i=1}^{n} k_{ij} = \text{efectivo total de la columna } j.$$

El método buscado para el análisis factorial de correspondencias simple deberá ser simétrico con relación a las líneas y columnas de *K* (para estudiar las relaciones en el interior de los conjuntos *I* y *J*) y deberá permitir comparar las distribuciones de frecuencias de las dos características (para estudiar las relaciones entre los conjuntos *I* y *J*).

En R^p tomaremos la nube de *n* puntos *i* (*n* filas de la tabla de perfiles de las variables *i*) cuyas coordenadas son $\frac{k_{i1}}{k_{i.}}, \frac{k_{i2}}{k_{i.}}, \ldots, \frac{k_{ip}}{k_{i.}}$ $i = 1, \ldots, n$

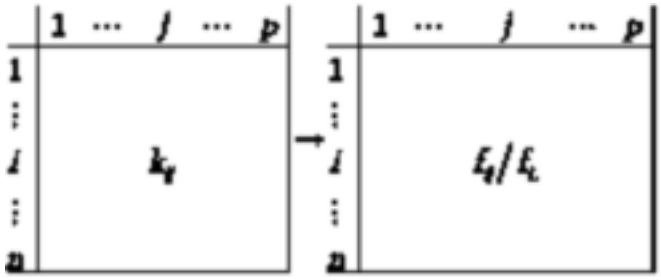
En R^n se forma la nube de *p* puntos j (*p* columnas de la tabla de perfiles de las observaciones *j*) cuyas coordenadas son $\frac{k_{1j}}{k_{.j}}, \frac{k_{2j}}{k_{.j}}, \ldots, \frac{k_{nj}}{k_{.j}}$ $j = 1, \ldots, p$

Las transformaciones realizadas son idénticas en los dos espacios R^p y R^n. Sin embargo, ello va a llevar a transformaciones analíticas diferentes. Los nuevos datos en R^n no son la traspuesta de la matriz en R^p. Esto nos conduce a *realizar dos análisis factoriales diferentes, uno en cada espacio.* Pero es posible encontrar relaciones entre los factores que permitirán reducir los cálculos a una sola factorización facilitando además la interpretación.

A partir de ahora se trabajará con la *tabla de contingencia en frecuencias relativas* $f_{ij} = \frac{k_{ij}}{k}$ con $k = \sum_{i=1}^{n} \sum_{j=1}^{p} k_{ij}$

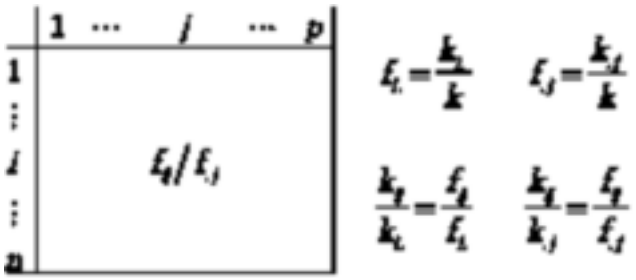
Tendremos el siguiente esquema:

Perfil de las líneas en R^p



fl

Perfil de las columnas en R^n



El análisis factorial de correspondencias trabaja con perfiles, pero no olvida las diferencias entre los efectivos de cada línea o columna, sino que les asigna un peso proporcional a su importancia en el total. En R^p cada punto *i* está afectado por un peso $f_{i.}$ y en R^n cada punto *j* está afectado por un peso $f_{.j}$ con lo que, de esta forma, se evita que al trabajar con perfiles se privilegie a las clases de efectivos pequeños.

El hecho de trabajar con perfiles, en vez de con los valores absolutos iniciales nos lleva a utilizar la distancia *Chi-cuadrado* (distancia entre distribuciones) en vez de la euclídea. Partiendo de la definición de distancia *Chi-cuadrado*, en el análisis de correspondencias la distancia entre los individuos (puntos fila) *i* e *i'* en R^p vendrá definida como:

$$d^2_{ii'} = \sum_{j=1}^{p} \frac{1}{k_{.j}/k}\left(\frac{k_{ij}}{k_{i.}} - \frac{k_{i'j}}{k_{i'.}}\right)^2 = \sum_{j=1}^{p} \frac{1}{k_{.j}/k}\left(\frac{k_{ij}/k}{k_{i.}/k} - \frac{k_{i'j}/k}{k_{i'.}/k}\right)^2 = \sum_{j=1}^{p} \frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$$

De forma similar, en el análisis de correspondencias la distancia entre las variables (puntos columna) *j* y *j'* en R^n vendrá definida como:

$$d^2_{jj'} = \sum_{i=1}^{n} \frac{1}{k_{i\cdot}/k} \left(\frac{k_{ij}}{k_{\cdot j}} - \frac{k_{ij'}}{k_{\cdot j'}} \right)^2 = \sum_{i=1}^{n} \frac{1}{k_{i\cdot}/k} \left(\frac{k_{ij}/k}{k_{\cdot j}/k} - \frac{k_{ij'}/k}{k_{\cdot j'}/k} \right)^2 = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2$$

El uso de la distancia *Chi-cuadrado* estabiliza los datos, hasta el punto de que, por el principio de la equivalencia distribucional, dos líneas (filas o columnas) con el mismo perfil pueden ser sustituidas por una sola afectada por una masa igual a la suma de las masas, sin que se alteren las distancias entre los demás pares de puntos en R^p o R^n.

Como el análisis es simétrico para filas y columnas, en el análisis factorial de correspondencias suele elegirse para columnas la dimensión más pequeña ($p<n$).

Para el análisis en R^p el objetivo es obtener una representación simplificada de los puntos fila cuyas coordenadas son $f_{ij}/f_{i\cdot}$, $j=1,...,p$. Estos puntos están afectados de un peso o masa $f_{i\cdot}$ y la distancia entre ellos se mide a través de la distancia *Chi-cuadrado*. Las coordenadas del centro de gravedad de la nube de puntos son $g_j = \sqrt{f_{\cdot j}}$.

La proyección de un punto sobre un nuevo eje de vector unitario u_α viene dada por el producto escalar del punto y el vector u_a, es decir:

$$F_a(i) = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) u_{aj}$$

Para hallar el factor *a*, se trata de buscar u_a que maximice la *inercia de la nube proyectada*, es decir, la suma de los cuadrados de las proyecciones cada una multiplicada por su peso ($\max \sum_{i=1}^{n} f_{i\cdot} F_a^2(i)$), problema equivalente a diagonalizar (vectores propios) la matriz *Z* de término general:

$$z_{jj'} = \sum_{i=1}^{n} f_{i\cdot} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \left(\frac{f_{ij'}}{f_{i\cdot}\sqrt{f_{\cdot j'}}} - \sqrt{f_{\cdot j'}} \right) = \sum_{i=1}^{n} \left(\frac{f_{ij} - f_{i\cdot}f_{\cdot j}}{\sqrt{f_{i\cdot}}\sqrt{f_{\cdot j}}} \right) \left(\frac{f_{ij'} - f_{i\cdot}f_{\cdot j'}}{\sqrt{f_{i\cdot}}\sqrt{f_{\cdot j'}}} \right)$$

Esta matriz se puede expresar como $Z=X'X$ siendo X la matriz de término general $x_{ij}=\frac{f_{ij}-f_{i.}f_{.j}}{\sqrt{f_{i.}}\sqrt{f_{.j}}}$.

Por lo tanto, el análisis factorial de correspondencias relativo a la tabla inicial k_{ij} es equivalente al análisis en componentes principales para la matriz de término general x_{ij}. De todas formas, se pueden realizar algunas simplificaciones, basadas en el hecho de que el vector u_p director del eje p que tiene de coordenadas ($f_{.1}$, $f_{.2}$,..., $f_{.p}$) es un vector propio de $Z = X'X$ asociado al valor propio 0.

Tenemos que todos los vectores propios de $Z=X'X$, "a p son también vectores propios de $S=X^{*'}X^{*}$ siendo $x_{ij}^{*}=\frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}}$.

El vector u_p es también vector propio de S, pero asociado al valor propio 1, por lo que el análisis puede realizarse sobre la tabla X^* no centrada. Esto conlleva que la proyección del punto i sobre el eje α (*coordenada factorial de la fila i*) vale:

$$F_{\alpha}(i)=\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}-\sqrt{f_{.j}}\right)u_{\alpha j}=\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}\right)u_{\alpha j}$$

Para el análisis en R^n tendremos que las coordenadas de los puntos j serán $f_{ij}/f_{.j}$, su peso será $f_{.j}$, el centro de gravedad G tendrá de coordenadas y la proyección de un punto j sobre el eje a cuyo vector director es v_a es:

$$G_{\alpha}(j)=\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}}-\sqrt{f_{i.}}\right)v_{\alpha i}$$

Para hallar el factor a, se trata de buscar v_a que maximice la inercia de la nube proyectada, es decir, la suma de los cuadrados de las proyecciones cada una multiplicada por su peso ($\max\sum_{j=1}^{p}f_{.j}G_{\alpha}^{2}(j)$), problema equivalente a diagonalizar (vectores propios) la matriz W de término general:

$$w_{ij}=\sum_{k=1}^{p}f_{.k}\left(\frac{f_{ik}}{f_{i.}\sqrt{f_{.k}}}-\sqrt{f_{.k}}\right)\left(\frac{f_{jk}}{f_{j.}\sqrt{f_{.k}}}-\sqrt{f_{.k}}\right)=\sum_{k=1}^{p}\left(\frac{f_{ik}-f_{i.}f_{.k}}{\sqrt{f_{i.}}\sqrt{f_{.k}}}\right)\left(\frac{f_{jk}-f_{j.}f_{.k}}{\sqrt{f_{j.}}\sqrt{f_{.k}}}\right)$$

Además, el vector v_p director del eje *p* de coordenadas ($f_1, f_2, ..., f_n$.) es un vector propio de *W=XX'* asociado al valor propio 0, y todos los vectores propios v_a de *W=XX'* "a *p* son también vectores propios de *W*=X*X*'* siendo v_p el vector propio asociado al valor propio 1. Esto conlleva a que la proyección del punto *j* sobre el eje *a* (*coordenada factorial de la columna j*) toma la expresión:

$$G_a(j)=\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}-\sqrt{f_{i\cdot}}\right)v_{ia}=\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}\right)v_{ia}$$

Pero como los valores propios 1_a no nulos de las matrices *X'X* y *XX'* son los mismos y los vectores propios u_a de *X'X* y v_a de *XX'* están relacionados, *es posible representar simultáneamente los puntos línea y los puntos columna sobre los mismos gráficos*, lo que favorece la interpretación de los resultados. Tenemos:

- La proyección de los puntos *j* sobre el eje α puede expresarse en función de la proyección de los puntos *i* (utilizando que $v_{ia}=\frac{1}{\sqrt{\lambda_a}}F_a(i)\sqrt{f_{i\cdot}}$) como sigue:

$$G_a(j)=\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}}\right)v_{ia}=\frac{1}{\sqrt{\lambda_a}}\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}}\right)F_a(i)\sqrt{f_{i\cdot}}=\frac{1}{\sqrt{\lambda_a}}\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{\cdot j}}\right)F_a(i)$$

- La proyección de los puntos *i* sobre el eje α puede expresarse en función de la proyección de los puntos *j* (utilizando que $u_{ja}=\frac{1}{\sqrt{\lambda_a}}G_a(j)\sqrt{f_{\cdot j}}$) como sigue:

$$F_a(i)=\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}\right)u_{ja}=\frac{1}{\sqrt{\lambda_a}}\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}\right)G_a(j)\sqrt{f_{\cdot j}}=\frac{1}{\sqrt{\lambda_a}}\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i\cdot}}\right)G_a(j)$$

Según las expresiones anteriores resultan las relaciones siguientes:

- La proyección de un punto *i* sobre el eje *a*, $F_a(i)$, es el baricentro (salvo el coeficiente $1/\sqrt{1_a}$) de las proyecciones de los puntos *j* sobre el mismo eje, cada punto afectado del peso *fij/fi.* que es su importancia relativa en *i*.

- La proyección de un punto *j* sobre el eje *a*, $G_a(j)$, es el baricentro (salvo el coeficiente $1/\lambda_a$) de las proyecciones de los puntos *i* sobre el mismo eje, cada punto afectado del peso *fij/f.j* que es su importancia relativa en *j*.

Las relaciones anteriores, llamadas *relaciones baricéntricas*, *permiten pasar de un espacio a otro y representar simultáneamente sobre el mismo plano los puntos fila y columna, permitiendo así clarificar las relaciones entre filas y columnas*.

En cuanto a las *contribuciones*, centraremos la atención en las modalidades que más claramente se asocian a un factor. Éstas son normalmente las que ocupan un lugar próximo al eje que representa el factor y que además están lejanas del origen de coordenadas factoriales. Son las modalidades que más inercia tienen las que definen el factor, y en esto interviene, además de las coordenadas factoriales de las modalidades, su masa. Si se denomina contribución absoluta de todas las modalidades de *I* al factor *a* a la expresión:

$$\lambda_a = C_a(I) = \sum_{i=1}^{n} C_a(i) = \sum_{i=1}^{n} f_i \cdot F_a^2(i), \quad i \in I$$

se tiene que $C_a(i) = f_i \cdot F_a^2(i)$ es la *contribución de la fila i a la inercia del eje a*.

De forma análoga, si se denomina contribución absoluta de todas las modalidades de *J* al factor *a* a:

$$\lambda_a = C_a(J) = \sum_{j=1}^{p} C_a(j) = \sum_{j=1}^{p} f_j \cdot G_a^2(j), \quad j \in J$$

se tiene que $C_a(j) = f_j \cdot G_a^2(j)$ es la *contribución de la fila i a la inercia del eje a*.

Las contribuciones absolutas no representan adecuadamente la importancia de un punto en la construcción de un eje factorial. Es necesario acudir a las *contribuciones relativas de las modalidades a los factores*.

Se denomina *contribución relativa de la modalidad i al factor a* al valor:

$$R(i/a) = \frac{C_a(i)}{C_a(I)} = \frac{f_i \cdot F_a^2(i)}{\sum_{i=1}^{n} f_i \cdot F_a^2(i)} = \frac{f_i \cdot F_a^2(i)}{\lambda_a}$$

Se denomina *contribución relativa de la modalidad j al factor* a al valor:

$$R(j/\alpha)=\frac{C_\alpha(j)}{C_\alpha(J)}=\frac{f_j G_\alpha^2(j)}{\sum_{j=1}^{p} f_j G_\alpha^2(j)}=\frac{f_j G_\alpha^2(j)}{\lambda_\alpha}$$

La suma de las contribuciones relativas de las filas es la unidad y lo mismo ocurre con las contribuciones relativas de las columnas.

Pero también es posible evaluar cómo está representada una fila o una columna por los distintos factores a través de las contribuciones relativas de lo factores sobre las modalidades.

Se denomina *contribución relativa del factor* a *sobre la modalidad de fila i* al valor:

$$R(\alpha/i)=\frac{C_\alpha(i)}{C(i)}=\frac{f_i F_\alpha^2(i)}{\sum_{\alpha=1}^{N} f_i F_\alpha^2(i)}=\frac{F_\alpha^2(i)}{\sum_{\alpha=1}^{N} F_\alpha^2(i)}$$

Se denomina *contribución relativa sobre el factor* a *de la modalidad de columna j* al valor:

$$R(\alpha/j)=\frac{C_\alpha(j)}{C(j)}=\frac{f_j G_\alpha^2(j)}{\sum_{\alpha=1}^{N} f_j G_\alpha^2(j)}=\frac{G_\alpha^2(j)}{\sum_{\alpha=1}^{N} G_\alpha^2(j)}$$

Para interpretar un factor es conveniente elegir un reducido número de modalidades cuya contribución a la inercia del factor sea fuerte. Para ello conviene buscar los puntos para los que la contribución relativa de modalidad a factor es elevada.

El examen de las contribuciones relativas de factor modalidad permite saber, para cada punto, si se encuentra alejado o no de la dirección del subespacio considerado.

Se entiende por calidad de la reconstrucción de una modalidad por medio de los primeros *m* ejes como la suma de las contribuciones relativas de esos *m* ejes sobre tal modalidad.

Puede concluirse que el análisis de correspondencias simple consiste en realizar un doble análisis de componentes principales sobre las nubes

de puntos fila y de puntos columna, ponderando cada punto de la nube por su masa y utilizando la métrica de la distancia *Chi-cuadrado* en el cálculo de las distancias.

Las nubes de puntos fila y puntos columna se representan en los planos de proyección formados por los primeros ejes factoriales tomados de dos en dos. Estos gráficos se interpretan de acuerdo a los valores propios, porcentajes de inercia y coeficientes relativos de las coordenadas factoriales.

Puede decirse que *dos variables son independientes si los perfiles de sus modalidades son idénticos a los perfiles medios*. En este caso, los puntos de las nubes se concentran alrededor del centro de gravedad adoptando una forma esférica. Mientras menor sea el valor de la inercia total, más concentrada estará la nube alrededor del centro de gravedad y más parecidos serán los perfiles al perfil medio (independencia). La *inercia* $\frac{1}{a}$ *de cada eje factorial* indica si existen o no direcciones privilegiadas en la nube. La existencia de grandes diferencias entre las tasas de inercia de unos ejes y otros indica direcciones privilegiadas en la tabla y nube no esférica (no independencia).

En general, se define el *porcentaje de inercia explicada por los k primeros ejes factoriales* como:

$$\tau_k = \frac{\sum_{a=1}^{k} \lambda_a}{\sum_{a=1}^{p} \lambda_a} = \frac{\sum_{a=1}^{k} \lambda_a}{\text{traza}(X'X)}$$

La *inercia total de la nube de puntos fila respecto a su centro de gravedad* es una medida de la dispersión de la nube y se calcula como la suma ponderada de las distancias entre los puntos fila y el centro de gravedad de la nube usando como ponderación la masa de cada punto fila y como métrica la distancia *Chi-cuadrado*. Su valor es:

$$I = \sum_{i=1}^{n} f_{i\cdot} d^2(i,G) = \sum_{i=1}^{n} f_{i\cdot} \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 = \sum_{i,j}^{np} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}}$$

De modo similar, la *inercia total de la nube de puntos columna respecto a su centro de gravedad* es:

$$J = \sum_{j=1}^{p} f_{\cdot j} d^2(j,G) = \sum_{j=1}^{p} f_{\cdot j} \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot} \right)^2 = \sum_{i,j}^{np} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}}$$

Se observa que las inercias totales de las nubes de puntos fila y columna coinciden y su valor es idéntico al del estadístico de la *Chi-cuadrado* para la independencia en una tabla de contingencia 2x2.

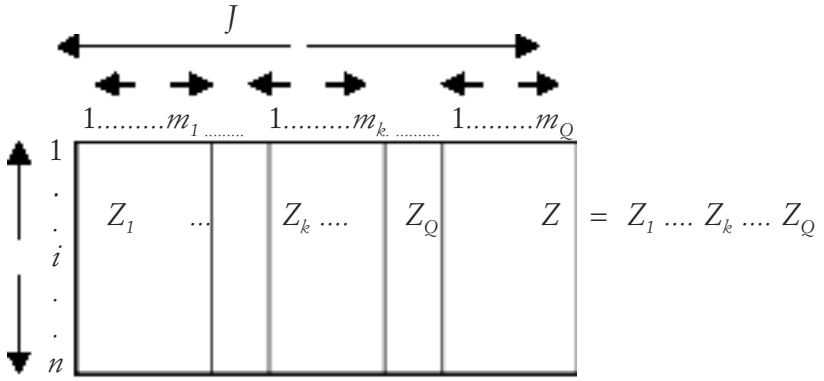
Se pueden ilustrar los planos factoriales obtenidos por análisis de correspondencias mediante informaciones (filas y columnas) que no han tomado parte en la construcción de tales planos que se denominan *filas y columnas suplementarias*.

Los elementos (filas y columnas) utilizadas para calcular los planos factoriales se denominan elementos activos y deben de formar un conjunto homogéneo y exhaustivo para que las distancias entre los elementos sea fácilmente interpretables, es decir, deben referirse a un mismo tema. Suelen analizarse como elementos suplementarios observaciones recogidas bajo condiciones poco claras o distintas de las del resto (elementos aberrantes o casos nuevos). También se tratan como suplementarios los elementos recogidos después de la realización del análisis.

11.3. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES ACM

El análisis de correspondencias simples o sencillamente análisis factorial de correspondencias se aplica cuando disponemos de dos caracteres o variables cualitativas, cada una de las cuales puede presentar varias modalidades o categorías. Pero el método es generalizable al caso de un número de variables o caracteres cualitativos mayor de dos, en cuyo caso estamos ante el *análisis de correspondencias múltiples*. Esta técnica permite describir grandes tablas lógicas con ceros y unos, como por ejemplo las que resultan de la codificación de una encuesta. Las filas de estas tablas suelen ser individuos u observaciones y las columnas son las modalidades de las variables nominales (modalidades de respuesta a cada una de las preguntas de una encuesta). El análisis de correspondencias múltiples puede considerarse como un análisis de correspondencias simples aplicado a una tabla disyuntiva completa, en lugar de a una tabla de contingencia.

En el *análisis de correspondencias múltiples* se ordenan los datos iniciales en una tabla *Z* denominada *tabla disyuntiva completa* que consta de un conjunto de individuos $I=1,...,i,...n$ (en filas), un conjunto de variables o caracteres cualitativos $J_1,...,J_k...J_Q$ (en columnas) y un conjunto de modalidades excluyentes $1,...,m_k$ para cada carácter cualitativo. El número total de modalidades será entonces $J=\sum_{k=1}^{Q} m_k$. La tabla disyuntiva completa *Z* de dimensión *IxJ* tiene el siguiente aspecto:



El elemento z_{ij} de la tabla toma el valor 0 ó 1 según que el individuo *i* haya elegido (esté afectado por) la modalidad *j* o no.

Por lo tanto, cada rectángulo de la tabla disyuntiva completa puede considerarse, aunque no lo sea, como una tabla de contingencia cuyos elementos son 0 ó 1. La tabla disyuntiva completa *Z* consta entonces de *Q* subtablas yuxtapuestas, con la finalidad de obtener una representación simultánea de todas las modalidades (columnas) de todos los individuos (filas). Si las modalidades son excluyentes, cada subtabla tiene un único 1 en cada una de sus filas.

A continuación se presenta un ejemplo de tabla disyuntiva completa *Z* generada a partir de una tabla inicial de datos *T*.

$$T=\begin{pmatrix}2&2&4\\2&1&3\\3&1&2\\1&2&4\\1&2&3\\3&2&3\\3&1&1\\1&1&1\end{pmatrix}\Rightarrow Z=\left(\begin{array}{ccc|cc|cccc}0&1&0&0&1&0&0&0&1\\0&1&0&1&0&0&0&1&0\\0&0&1&1&0&0&1&0&0\\1&0&0&0&1&0&0&0&1\\1&0&0&0&1&0&0&1&0\\0&0&1&0&1&0&0&1&0\\0&0&1&1&0&1&0&0&0\\1&0&0&1&0&1&0&0&0\end{array}\right)$$

Si conservamos la notación que hemos manejado hasta ahora tenemos:

$$z_{ij} = k_{ij} = 0 \text{ ó } 1.$$

$k_{i.} = \sum_j k_{ij}$ = Q = número de modalidades (cada subtabla tiene un único 1 en cada fila).

$k_{.j} = \sum_i k_{ij}$ = número de individuos que poseen la modalidad *j*.

$f_{ij}/f_{i.} = k_{ij}/k_{i.} = 1/Q$ = inverso del número de modalidades (0 si el individuo no elige *j*).

Para obtener los factores es necesario diagonalizar la matriz $V=D^{-1}B/Q$ donde $B=Z'Z$ *es la tabla de contingencia de Burtz*, matriz simétrica formada por Q^2 bloques, de modo que sus bloques de la diagonal Z'_kZ_k son tablas diagonales que cruzan una variable con ella misma, siendo los elementos de la diagonal los efectivos de cada modalidad $k_{.j}$. Los bloques fuera de la diagonal son tablas de contingencia obtenidas cruzando las características de dos en dos Z'_kZ_k cuyos elementos son las frecuencias de asociación de las dos modalidades correspondientes. La matriz *D* es una matriz diagonal cuyos elementos diagonales son los de la matriz de Burtz, siendo nulos el resto de los elementos. La matriz *Z* es la tabla disyuntiva completa. El aspecto de la tabla de Burt es el siguiente:

	J_1	J_2	...	J_Q
J_1	$0^{\cdot}\cdot.0$	C_{12}	⋯	C_{1Q}
J_2	C_{21}	$0^{\cdot}\cdot.0$	⋯	C_{2Q}
⋮	⋮	⋮	⋱	⋮
J_Q	C_{Q1}	C_{Q2}	⋯	$0^{\cdot}\cdot.0$

Las fórmulas de transición que *permiten representar simultáneamente los puntos línea y los puntos columna sobre los mismos gráficos relacionando así los resultados en los dos subespacios* tomarán ahora las siguientes expresiones:

$$F_a(i) = \frac{1}{\sqrt{\lambda_a}} \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.}} \right) G_a(j) = \frac{1}{\sqrt{\lambda_a}} \frac{1}{Q} \sum_{j=1}^{p} k_{ij} G_a(j)$$

$$G_a(j) = \frac{1}{\sqrt{\lambda_a}} \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{.j}} \right) F_a(i) = \frac{1}{\sqrt{\lambda_a}} \frac{1}{k_{.j}} \sum_{i=1}^{n} k_{ij} F_a(i)$$

Si tenemos en cuenta que $k_{ij}=1$ cuando el individuo *i* posee la modalidad *j* y cero cuando no, la *proyección de un punto individuo i sobre el eje a*, $F_a(i)$, es el baricentro (salvo un coeficiente de dilatación $1/\sqrt{\lambda_a}$) de las pro-

yecciones de los puntos modalidades sobre el eje, $G_a(j)$. Todas las modalidades están afectadas del mismo peso $1/Q$. Análogamente, la *proyección de un punto modalidad j sobre el eje a, $G_a(j)$*, es el baricentro (salvo un coeficiente de dilatación $1/\lambda_a$) de las proyecciones de los puntos individuos que poseen esa modalidad sobre el eje, $F_a(i)$, todos ellos afectados del mismo peso $k_{.j}$.

El *centro de gravedad de la nube de puntos variables N(j)* en análisis factorial de correspondencias (*ACM*) es $f_{i.}$, que en este caso puede equipararse a una distribución uniforme $1/n$, ya que

$$k_{i.} = \sum_j k_{ij} = Q \Rightarrow \sum_i k_{i.} = nQ \Rightarrow f_{i.} = 1/n$$

El *centro de gravedad de las modalidades de cada variable*, cada una ponderada por su peso, es el mismo que el de la nube de modalidades *N(J)*, es decir, $1/n$, ya que el centro de gravedad de la subtabla IxJ_k se obtiene a partir de su distribución marginal. Como sólo recoge una variable, la suma de cada línea es 1 y el total de la tabla es n, de dónde $f_{i.}=1/n$.

Como el análisis factorial de correspondencias es centrado y el centro de gravedad de las modalidades de una variable coincide con el del conjunto *J*, y con el origen, las modalidades de cada variable están centradas en torno al origen, no pudiendo tener todas el mismo signo.

Al igual que en cualquier análisis factorial de correspondencias, se calculan las *ayudas a la interpretación para cada fila y columna*, definiendo la contribución de una variable J_k al factor *a*, como la suma de las contribuciones de las modalidades de la variable:

$$CTA_a(J_k) = \sum_{j \in J_k} CTA_a(j)$$

La parte de inercia debida a una modalidad *j* es mayor cuanto menor sea el efectivo de esa modalidad. Si *G* representa el centro de gravedad, la *inercia debida a la modalidad j* viene dada por:

$$I(j) = f_{.j} \cdot d^2(G, j) = f_{.j} \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \right)^2 = \frac{k_j}{nQ} \sum_{i=1}^{n} \left(\frac{k_{ij}/nQ}{k_{i.} \cdot 1/n} - 1/\sqrt{n} \right)^2 = \frac{1}{Q}\left(1 - \frac{k_j}{n}\right)$$

Por lo tanto, es aconsejable eliminar las modalidades elegidas muy pocas veces, construyendo otra modalidad uniéndola a la más próxima.

La parte de *inercia debida a una variable* es función creciente del número de modalidades de respuesta que tiene, ya que la inercia de una variable es la suma de las inercias de sus modalidades:

$$I(J_i)=\sum_{j\in J_i} I(j)=\sum_{j\in J_i}\frac{1}{Q}\left(1-\frac{k_j}{n}\right)=\frac{1}{Q}(m_i-1)$$

Si una variable tiene un número de modalidades demasiado grande, al igual que en el caso de que su efectivo sea muy pequeño, conviene reagrupar las modalidades en un número que sea razonable y mantenga el sentido, para evitar así influencias extremas.

La *inercia total* es la suma de las inercias de todas las modalidades:

$$I=\sum_i I(J_i)=\sum_i\frac{1}{Q}(m_i-1)=\frac{J}{Q}-1$$

J/Q es el número medio de modalidades por variable cualitativa o carácter. En consecuencia, la inercia total sólo depende del número de modalidades y del de preguntas.

Si el número de variables es dos, y cada una tiene dos modalidades, los resultados se pueden analizar tanto por análisis factorial de correspondencias (AFC) como por análisis de correspondencias múltiples (ACM). En el primer caso obtendríamos un único factor que recoge el 100% de la inercia total. Esta inercia dependerá del grado de relación que exista entre las modalidades, de modo que, si están poco relacionadas, la inercia será próxima a cero, y si están muy relacionadas, la inercia tenderá a un valor alto.

Si la misma información la analizamos mediante análisis de correspondencias múltiples, obtendremos siempre la misma inercia (*J*/*Q*-1=1), pero obtendremos dos ejes. En el caso en que existe mucha relación entre las variables, el primer eje recogerá gran parte de la inercia (casi 1) y el segundo muy poca, mientras que en el caso de total independencia entre las dos variables ambos factores recogerán la misma cantidad de inercia, es decir, 1/2 cada uno.

El análisis en correspondencias múltiples pone en evidencia tipos de individuos que tienen perfiles semejantes en cuanto a los atributos que los describen. Teniendo en cuenta las distancias entre los elementos de la tabla disyuntiva completa y las relaciones baricéntricas puede decirse que dos individuos son próximos si presentan globalmente las mismas modalida-

des. La proximidad entre modalidades de variables en términos de asociación va referida a los puntos medios de los individuos que las presentan. Las modalidades son próximas porque les corresponden globalmente los mismos individuos o individuos semejantes. En cuanto a la proximidad entre modalidades de una misma variable, hay que tener en cuenta que las modalidades de una misma variable se excluyen. Su proximidad se interpreta en términos de semejanza entre los grupos de individuos que las presentan, respecto del resto de las variables activas del análisis.

A partir de la descomposición de la inercia de la nube de las modalidades, se calcula la contribución de una variable al factor *a* sumando las contribuciones de sus modalidades a ese factor. Así, además de las modalidades responsables de los ejes factoriales, se encuentran variables que han participado en la definición del factor. Se obtiene así un indicador de la relación entre la variable y el factor.

Las reglas de interpretación de los resultados (contribuciones) relativos a los elementos de un análisis en correspondencias múltiples son prácticamente iguales que las de un análisis de correspondencias simples, siendo posible calcular la contribución y la calidad de la representación de cada modalidad y de cada individuo.

EJERCICIO 13-1. El gobierno pretende lanzar una campaña a nivel nacional sobre la responsabilidad de que los hijos lleguen puntuales a clase El fichero PISA2003 contiene información sobre las Comunidades Autónomas que participaron en la prueba:

Variable CCAA

7241 – Otras Comunidades Autónomas
7242 – Castilla y León
7243 – Cataluña
7244 – País Vasco

Variable LATE: Respuestas a la pregunta «En las últimas dos semanas que asististe al colegio ¿cuántas veces llegaste tarde? Respuestas

1 – Nunca
2 – 1 o 2 veces
3 – 3 o 4 veces
4 – 5 o más veces
9 – Dato perdido

Compruebe mediante un análisis de correspondencias simples si el gobierno debería hacer más énfasis en la campaña en alguna CC. AA. o si por el contrario el porcentaje de alumnos que llegan tarde es semejante en todas las CC. AA.

Para realizar este cálculo se acude a XLSTAT → Análisis de Datos → Análisis Factorial de Correspondencias (Figura 13-1).

FIGURA 13-1

Una vez seleccionadas las opciones contenidas en las pestañas los principales resultados obtenidos son los siguientes:

Tabla de contingencia

	LATE-1	*LATE-2*	*LATE-3*	*LATE-4*	*LATE-9*
CCAA-7241	2134	937	249	271	32
CCAA-7242	941	346	89	92	7
CCAA-7243	868	420	119	93	7
CCAA-7244	2380	897	249	253	11

Inercia por casilla

	LATE-1	*LATE-2*	*LATE-3*	*LATE-4*	*LATE-9*
CCAA-7241	0,00021	0,00010	0,00000	0,00022	0,00071
CCAA-7242	0,00021	0,00014	0,00012	0,00007	0,00001
CCAA-7243	0,00025	0,00047	0,00026	0,00009	0,00002
CCAA-7244	0,00023	0,00026	0,00003	0,00001	0,00044

Prueba de independencia entre filas y columnas

Chi-cuadrado ajustado (Valor observado)	40,195
Chi-cuadrado ajustado (Valor crítico)	21,026
GDL	12
p-valor	< 0,0001
alfa	0,05

Como el p-valor computado es menor que el nivel de significación alfa=0,05, se debe rechazar la hipótesis nula H0, y aceptar la hipótesis alternativa Ha.

Valores propios y porcentajes de inercia

	F1	*F2*	*F3*
Valor propio	0,003	0,001	0,000
Las filas dependen de las columnas (%)	71,895	25,200	2,905
% acumulado	71,895	97,095	100,000

Pesos, distancias y distancias cuadradas al origen, inercias e inercias relativas (filas)

	Peso (relativo)	*Distancia*	*Distancia 2*	*Inercia*	*Inercia relativa*
CCAA-7241	0,349	0,060	0,004	0,00125	0,324
CCAA-7242	0,142	0,062	0,004	0,00055	0,142
CCAA-7243	0,145	0,087	0,008	0,00109	0,282
CCAA-7244	0,365	0,052	0,003	0,00098	0,253

Pesos, distancias y distancias cuadradas al origen, inercias e inercias relativas (columnas)

	Peso (relativo)	*Distancia*	*Distancia 2*	*Inercia*	*Inercia relativa*
LATE-1	0,608	0,038	0,001	0,001	0,233
LATE-2	0,250	0,062	0,004	0,001	0,252
LATE-3	0,068	0,078	0,006	0,000	0,106
LATE-4	0,068	0,076	0,006	0,000	0,102
LATE-9	0,005	0,466	0,217	0,001	0,307

Perfiles (filas)

	LATE-1	*LATE-2*	*LATE-3*	*LATE-4*	*LATE-9*	*Suma*
CCAA-7241	0,589	0,259	0,069	0,075	0,009	1
CCAA-7242	0,638	0,235	0,060	0,062	0,005	1
CCAA-7243	0,576	0,279	0,079	0,062	0,005	1
CCAA-7244	0,628	0,237	0,066	0,067	0,003	1
Media	0,608	0,252	0,068	0,066	0,005	1

Perfiles (columnas)

	LATE-1	*LATE-2*	*LATE-3*	*LATE-4*	*LATE-9*	*Media*
CCAA-7241	0,337	0,360	0,353	0,382	0,561	0,399
CCAA-7242	0,149	0,133	0,126	0,130	0,123	0,132
CCAA-7243	0,137	0,162	0,169	0,131	0,123	0,144
CCAA-7244	0,376	0,345	0,353	0,357	0,193	0,325
Suma	1,000	1,000	1,000	1,000	1,000	1,000

La tabla de perfiles por filas muestra las frecuencias relativas de cada casilla frente al total de su fila. Así, los individuos encuestados en el resto de España son mucho más numerosos que en el resto de CC. AA. por representar a una mayor cantidad de población. En la tabla de perfiles columna se observa como en el resto de España las categorías con más porcentaje son LATE-4 y LATE-9 mientras que en el País Vasco y Castilla y León la categoría más frecuente es LATE-1.

Distancias chi-cuadrado (filas)

	CCAA-7241	*CCAA-7242*	*CCAA-7243*	*CCAA-7244*
CCAA-7241	0	0,112	0,096	0,109
CCAA-7242	0,112	0	0,139	0,039
CCAA-7243	0,096	0,139	0	0,123
CCAA-7244	0,109	0,039	0,123	0

Distancias chi-cuadrado (columnas)

	LATE-1	*LATE-2*	*LATE-3*	*LATE-4*	*LATE-9*
LATE-1	0	0,100	0,112	0,098	0,492
LATE-2	0,100	0	0,032	0,091	0,436
LATE-3	0,112	0,032	0	0,111	0,458
LATE-4	0,098	0,091	0,111	0	0,408
LATE-9	0,492	0,436	0,458	0,408	0

Coordenadas principales (filas)

	F1	*F2*	*F3*
CCAA-7241	-0,053	0,029	-0,002
CCAA-7242	0,056	0,014	0,023
CCAA-7243	-0,055	-0,067	0,005
CCAA-7244	0,050	-0,007	-0,009

Coordenadas principales (columnas)

	F1	*F2*	*F3*
LATE-1	0,038	0,007	0,003
LATE-2	-0,057	-0,024	0,004
LATE-3	-0,057	-0,051	-0,013
LATE-4	-0,039	0,057	-0,032
LATE-9	-0,373	0,270	0,064

Las coordenadas de filas y columnas en el plano definido por dos factores serán los parámetros que se utilizan para la confección del mapa de categorías por filas y columnas.

Tal y como puede observarse a simple vista por la cercanía entre categorías Castilla y León y el País Vasco son las CC. AA. donde los alumnos son más puntuales. A continuación estaría Cataluña que se asocia a las respuestas de entre 1 y 4 veces tarde en dos semanas. En el resto de CC. AA. están los alumnos que declaran en mayor frecuencia haber llegado 5 o más veces tarde en las dos semanas previas a la encuesta. La categoría de valores perdidos no se asocia a ninguna CC. AA. en especial si bien podríamos trazar una recta que uniera la categoría LATE-9 con CCAA-7241 y CCAA-7244.

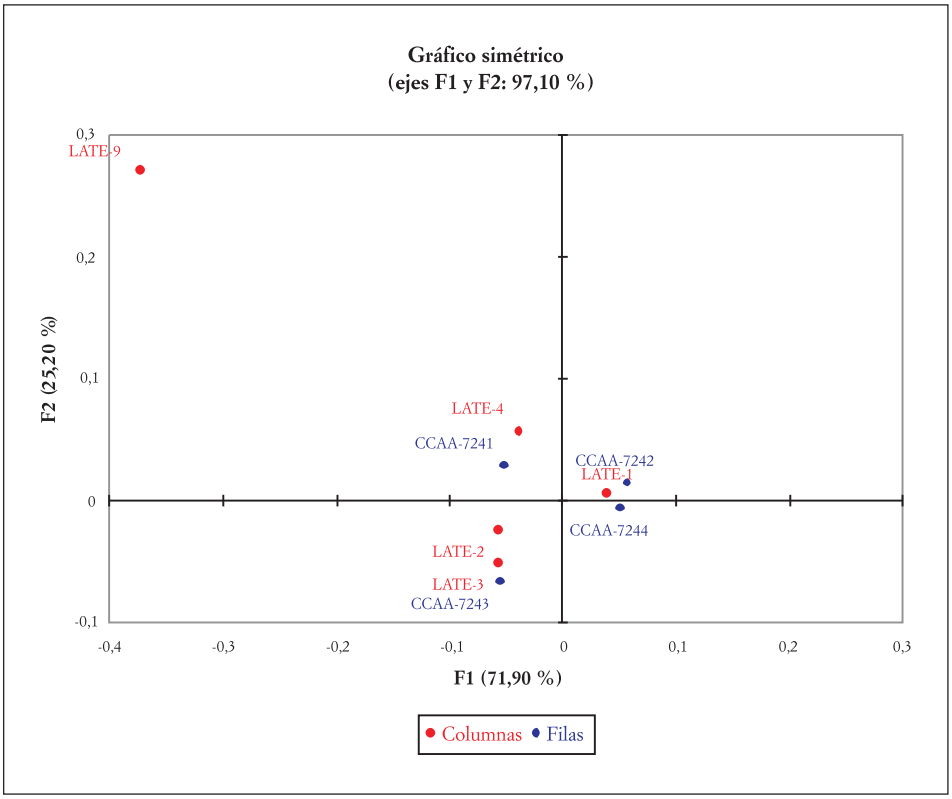


GRÁFICO SIMÉTRICO O MAPA DE CATEGORÍAS

Vistos estos resultados recomendaríamos ejercer una mayor difusión de la campaña en el resto de España.

EJERCICIO 13-2. Supongo ahora que se desea añadir al análisis si la influencia del tipo de familia influye en el retraso de llegada a la escuela. El fichero PISA2003 contiene información sobre las Comunidades Autónomas que participaron en la prueba:

Variable FAMSTRUCT

1 – Monoparental
2 – Nuclear
3 – Mixta o recompuesta
4 – Otro tipo
9 – No contesta

Variable LATE: Respuestas a la pregunta «En las últimas dos semanas que asististe al colegio ¿cuántas veces llegaste tarde? Respuestas

1 – Nunca
2 – 1 o 2 veces
3 – 3 o 4 veces
4 – 5 o más veces
9 – No contesta

Compruebe mediante un análisis de correspondencias simples si el gobierno debería destinar más recursos a alguno de los tipos de familia.

Para realizar este cálculo se acude a XLSTAT → Análisis de Datos → Análisis de Correspondencias Múltiples (Figura 13-2).

FIGURA 13-2

Una vez seleccionados los datos en las pestañas resultados y gráficos seleccionamos el resto de opciones deseadas y ejecutamos el análisis. Los resultados obtenidos son los siguientes.

Estadísticas Simples

Variable	*Categorías*	*Frecuencias*	*%*
CCAA	7241	3623	34,853
	7242	1475	14,190
	7243	1507	14,497
	7244	3790	36,460
LATE	1	6323	60,827
	2	2600	25,012
	3	706	6,792
	4	709	6,821
	9	57	0,548
FAMSTRUCT	1	1335	12,843
	2	8538	82,136
	3	250	2,405
	4	209	2,011
	9	63	0,606

Valores propios y porcentajes de inercia

	F1	*F2*	*F3*	*F4*	*F5*
Valor propio	0,379	0,353	0,346	0,338	0,336
Inercia (%)	10,335	9,627	9,436	9,223	9,162
% acumulado	10,335	19,962	29,398	38,620	47,782
Inercia ajustada	0,005	0,001	0,000	0,000	0,000
Inercia ajustada (%)	65,193	12,111	5,017	0,730	0,212
% acumulado	65,193	77,304	82,321	83,051	83,262

El gráfico de correspondencias múltiples muestra resultados interesantes. En primer lugar la familia nuclear aparece claramente asociada a los alumnos que declaran no llegar tarde nunca. La familia mixta aparece cercana a la categoría LATE-2 y la familia monoparental a la categoría LATE-3. Los alumnos que pertenecen a otros tipos de familia diferentes de las anteriores tienen una asociación más clara a la categoría que más veces llega tarde a clase.

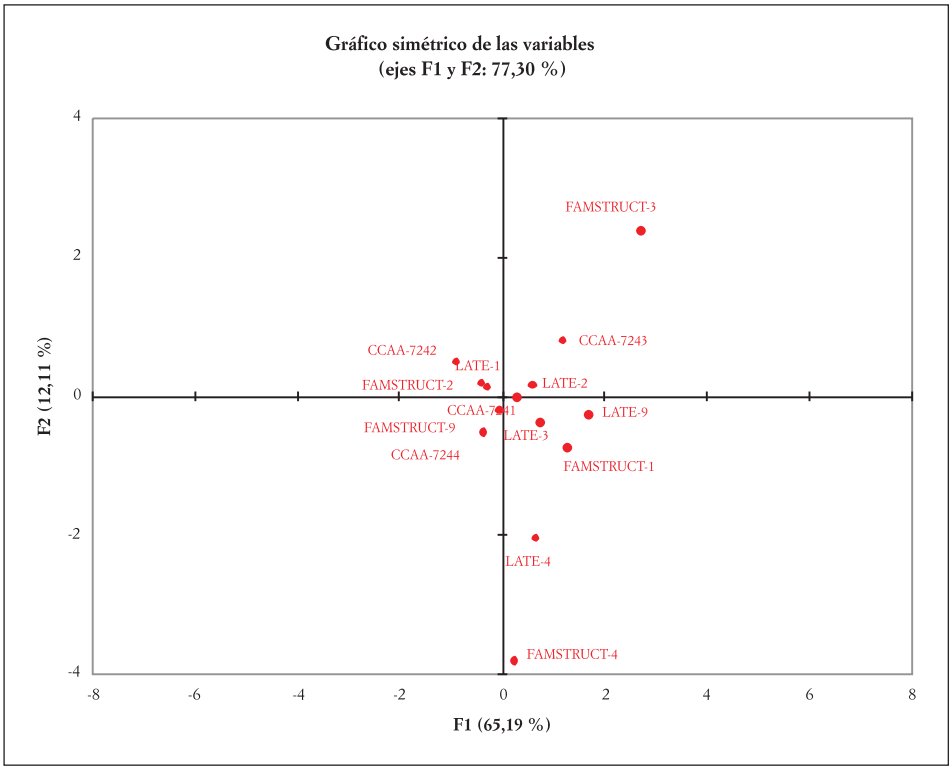


GRÁFICO SIMÉTRICO DE LAS VARIABLES

En definitiva, del análisis llevado a cabo concluiríamos que los alumnos con menor probabilidad de llegar tarde a clase serían los alumnos castellano-leonenses en familias nucleares mientras que los alumnos del resto de España en otro tipo de familias y en familias monoparentales serían los que más probabilidades tendrían de llegar tarde a clase y por tanto sobre los que convendría aplicar un mayor esfuerzo en la difusión de la información.

BIBLIOGRAFÍA

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. y Tatham, R. L. (2006). Multivariate Data Analysis. Sixth Edition. Pearson, Prentice Hall. New Jersey.

Pérez López, C. (2005). Métodos estadísticos avanzados con SPSS. Thomson Paraninfo.

Pérez López, C. (2005). Técnicas estadísticas con SPSS12. Aplicaciones al análisis de datos. Pearson Alambra.

Pérez López, C. (2004). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS. Pearson Alhambra.

CAPÍTULO XIV
ESCALADO MULTIDIMENSIONAL

CÉSAR PÉREZ LÓPEZ
DANIEL SANTÍN GONZÁLEZ

14.1. LA TÉCNICA DEL ESCALAMIENTO MULTIDIMENSIONAL

Podríamos definir el *escalamiento multidimensional* como un conjunto de técnicas que identifican dimensiones subyacentes a las evaluaciones de objetos hechas por los individuos o familias encuestadas cuyo propósito es transformar sus juicios en posiciones espaciales. El escalamiento multidimensional trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos. Esto se logra asignando las observaciones a posiciones específicas en un espacio conceptual (normalmente de dos o tres dimensiones) de modo que las distancias entre los puntos en el espacio concuerden al máximo con las disimilaridades o preferencias dadas. El objetivo del escalamiento multidimensional es transformar los juicios de similitud o preferencias llevados a cabo por una serie de individuos en distancias susceptibles de ser representadas en un espacio multidimensional.

Por ejemplo una empresa preocupada por la opinión de las familias en el tratamiento que ésta hace de determinados valores: solidaridad, conciliación, honradez, ayuda al tercer mundo, etc. puede estar interesada en realizar un escalamiento multidimensional para saber cual es su posición respecto a otras empresas de su mismo ramo de negocio. En función de las medidas que tome para favorecer el cambio de opinión la realización de un nuevo análisis transcurrido un tiempo nos indicará cuál es la nueva posición de la empresa.

El escalamiento multidimensional se clasifica dentro de los métodos de interdependencia y es un procedimiento que permite al investigador social determinar la imagen relativa percibida de un conjunto de objetos (valores, empresas, opiniones u otros objetos sobre los que los individuos desarrollan percepciones). Es decir, el aspecto característico de este procedimiento es que proporciona una representación gráfica en un espacio geométrico de pocas dimensiones (*mapa perceptual*) que permite comprender cómo los individuos perciben objetos y qué esquemas, generalmente ocultos, están detrás de esa percepción (en este sentido también se puede considerar el escalamiento multidimensional como una técnica de reducción de la dimensión).

En estos espacios, los objetos adoptan la forma de puntos y la proximidad entre ellos refleja la analogía existente entre los mismos. La interpretación de las dimensiones depende del conocimiento que se tenga acerca de esos estímulos por parte del investigador y se realiza de forma similar a como se haría con un análisis factorial clásico o un análisis de correspondencias.

Respecto a la elección del tipo de datos, el investigador debe optar entre la obtención de datos de similitud o de preferencias. Los mapas perceptuales basados en similitudes representan el parecido entre los atributos de los objetos. Los mapas perceptuales basados en datos de preferencias reflejan qué objetos son preferidos. En lo referente a la elección del método de análisis, se pueden utilizar métodos no métricos y métricos. Los métodos no métricos, llamados así por el carácter no métrico de los datos de entrada (comúnmente generados mediante la ordenación de pares de objetos), resultan más flexibles al no asumir ningún tipo específico de relación entre la distancia calculada y la medida de similitud. Sin embargo, es más probable que resulten en soluciones degeneradas o no óptimas. Los métodos métricos se distinguen por el carácter métrico tanto de los datos de entrada como de los resultados. La métrica nos permite reforzar la relación entre la dimensionalidad de la solución final y los datos iniciales.

Podrían considerarse varios pasos para determinar la posición de cada objeto en el espacio perceptual de modo que los múltiples juicios de similitud expresados por los individuos entrevistados se reflejen lo más fielmente posible. Un primer paso sería la selección de una configuración inicial de los estímulos según la dimensionalidad inicial deseada. Un segundo paso sería el cálculo de las distancias entre los puntos representativos de los estímulos y comparación de las relaciones (observadas versus derivadas) mediante una medida de ajuste o *Stress* (que indica la proporción de varianza de los datos originales no recogida por el modelo de escalamiento multidimensional). Si el indicador de ajuste no alcanza un valor mínimo previamente fijado por el investigador, un tercer paso sería encontrar una nueva configuración para la que el indicador de ajuste sea mejor. En un cuarto paso, el programa realizará una evaluación de la nueva configuración y la ajustará hasta que se logre obtener un nivel satisfactorio de ajuste. Un quinto y último paso sería la reducción de la dimensionalidad de la configuración actual y repetición del proceso hasta lograr obtener aquella configuración que, con la menor dimensionalidad posible, presente un nivel de ajuste aceptable (queda reforzada la idea de considerar el escalamiento multidimensional como una técnica de reducción de la dimensión). El analista debe preocuparse de obtener varias soluciones con diferente número de dimensiones y elegir entre ellas sobre la base de tres criterios fundamentales: su nivel de ajuste a los datos, su interpretabilidad y su replicabilidad.

14.2. FORMALIZACIÓN DE LA TÉCNICA DE ESCALAMIENTO MULTIDIMENSIONAL

Ya sabemos que el escalamiento multidimensional es una técnica de análisis multivariante que permite representar las proximidades entre un conjunto de objetos o estímulos como distancias en un espacio de baja dimensionalidad (generalmente de 2 ó 3 dimensiones). De modo más formal y general, nos centraremos en el hecho de que el escalamiento multidimensional toma como entrada habitual una matriz cuadrada de proximidades, a la que llamaremos Δ (delta), de dimensiones (*n,n*), donde *n* es el número de estímulos. Cada elemento ∂_{ij} de Δ representa la proximidad entre los estímulos *i* y *j*. Para $n = 4$, la matriz Δ tendría los siguientes elementos:

$$\Delta = \begin{bmatrix} \partial_{11} & \partial_{12} & \partial_{13} & \partial_{14} \\ \partial_{21} & \partial_{22} & \partial_{23} & \partial_{24} \\ \partial_{31} & \partial_{32} & \partial_{33} & \partial_{34} \\ \partial_{41} & \partial_{42} & \partial_{43} & \partial_{44} \end{bmatrix}$$

A partir de esa matriz de proximidades, el análisis de escalado multidimensional nos proporciona como solución una matriz rectangular *X*, de tamaño *n* x *m*, donde *n* es, al igual que antes, el número de estímulos, y *m* es el número de dimensiones.

Cada valor x_{ij} de la matriz *X* corresponde a la coordenada del estímulo *i* en la dimensión *j*. En escalamiento multidimensional la dimensionalidad utilizada siempre es la menor posible (2 ó 3 dimensiones en la mayoría de los casos, siendo muy raras las soluciones de dimensionalidad superior a 4). La matriz *X* correspondiente a una solución en 2 dimensiones para los 4 estímulos anteriores tendría los siguientes elementos:

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix}$$

Cada fila de esta matriz [X_{i1}, X_{i2}]contiene las coordenadas del estímulo *i* en los ejes de coordenadas *X* e *Y* que delimitan el espacio bidimensional. A partir de la matriz *X* es posible situar los *n* estímulos en el espacio asignándoles los valores de coordenadas correspondientes. También es posible utilizar la matriz *X* para calcular las distancias entre

dos estímulos *i* y *j* cualesquiera aplicando la fórmula general de la distancia de Minkowski:

$$d_{ij}=\left(\sum_{a=1}^{n}\left(x_{ia}-x_{ja}\right)^{p}\right)^{\frac{1}{p}} \quad (1\leq p\leq\infty)$$

Cuando $p = 2$, la distancia anterior es la métrica euclídea.

La estimación de las distancias correspondientes a todos los estímulos nos proporciona una nueva matriz, que llamaremos *D*. En el caso de nuestro ejemplo, los elementos de la matriz serían los siguientes

$$D=\begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{31} & d_{32} & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{bmatrix}$$

La solución del el escalamiento multidimensional debe proporcionar la máxima correspondencia entre las proximidades entre estímulos proporcionadas en la matriz Δ y las distancias entre estímulos obtenidas en la matriz D.

14.3. MODELO DE ESCALAMIENTO MÉTRICO

La relación asumida entre los datos de entrada (las proximidades) y las distancias entre estímulos obtenidos como solución determinan la tipología de los modelos de escalamiento multidimensional. Las distancias son función de las proximidades mediante $d_{ij}=f(\delta_{ij})$.

Se denominan modelos de escalamiento métrico aquéllos en que la función *f* es una función lineal con pendiente positiva. Tendremos entonces que:

$$\delta_{ij}\rightarrow a+b\delta_{ij}=d_{ij} \quad b>0$$

En el procedimiento de escalamiento multidimensional métrico, a partir de una matriz $D(n\times n)$ de distancias entre *n* estímulos se puede derivar una matriz de productos escalares entre vectores. A su vez, es posible descomponer la matriz *B* de productos escalares en el producto *XX'*, donde $X(n\times m)$ es la matriz de coordenadas de los *n* estímulos en *m* dimensiones. Adicionalmente, se puede llevar a cabo una transformación de la matriz de proximida-

des $\Delta(n \times m)$ en una matriz de distancias que respete los axiomas de la función de distancia euclídea $(d_{ii} = d_{jj} = 0,\ d_{ij} = d_{ji}\ \text{y}\ d_{ij} \leq d_{ik} + d_{kj})$.

Los dos primeros axiomas son fáciles de cumplir, pero para que se cumpla el tercero hay que buscar un valor *c* que, sumado a las proximidades originales (δ_{ij}) nos proporcione las distancias $(d_{ij} = \delta_{ij} + c)$. El valor mínimo de *c* que satisface la desigualdad triangular $(d_{ij} \leq d_{ik} + d_{kj})$ para toda terna de estímulos (*i, j, k*) se define como:

$$c_{min} = \max_{i,j,k}(\delta_{ij} - \delta_{ik} - \delta_{kj})$$

Calculada la matriz $D(n \times n)$, es necesario transformarla en una matriz $B(n \times n)$ de productos escalares entre vectores, de modo que los elementos b_{ij} de esta nueva matriz se crean a partir de los elementos d_{ij} de *D* mediante la siguiente transformación:

$$b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right) \quad \text{con} \quad d_{i.}^2 = \frac{1}{n}\sum_{j}^{n} d_{ij}^2, \quad d_{.j}^2 = \frac{1}{n}\sum_{i}^{n} d_{ij}^2 \ \text{y}$$

$$d_{..}^2 = \frac{1}{n^2}\sum_{i}^{n}\sum_{j}^{n} d_{ij}^2$$

A continuación se calcula la matriz de coordenadas *X* tal que *B=XX'*. En ocasiones resulta interesante, una vez obtenida la matriz *X*, rotar la solución para mejorar la interpretabilidad del resultado. La rotación de los ejes no altera las distancias entre los estímulos, por lo que es posible multiplicar la matriz *X* por una matriz de transformación ortogonal $T(r \times r)$, tal que *TT' = I*, donde *I* es la matriz identidad.

La matriz $X^* = XT$ contiene las coordenadas de los estímulos en la nueva solución rotada. Esta matriz es equivalente a la matriz *X*, ya que si *B = XX' , B = X*X*'*. Esto es así porque *X*X*' = XT T'X' = XIX' =XX'*.

El procedimiento expuesto fue ideado por Torgerson y posteriormente derivó en procedimientos iterativos.

14.4. MODELO DE ESCALAMIENTO NO MÉTRICO

En los modelos de escalamiento no métrico se asume la relación entre los datos de entrada (las proximidades) y las distancias entre estímulos obtenidos como solución $d_{ij} = f(\theta_{ij})$ cumple que *f* es una función monótona creciente. En este caso la relación entre proximidades y distancia es:

$$\theta_{ij} < \theta_{kl} \Rightarrow d_{ij} \leq d_{kl}.$$

En el escalado multidimensional no métrico se comienza convirtiendo las proximidades en

rangos, de 1 a $\frac{(n(n-1))}{2}$.

A continuación se crea una matriz de coordenadas aleatorias $X(n \times n)$. Es decir, se sitúan los estímulos al azar en un espacio de *r* dimensiones (donde *r* es especificado por el usuario). A partir de esta matriz *X* inicial se calculan las distancias entre estímulos. Estas distancias se comparan luego con los rangos de las proximidades, transformándolas si es necesario para que sus rangos coincidan con éstos. A las distancias obtenidas tras estas transformaciones se las denomina pseudodistancias o disparidades $(\hat{d}_{ij})$

En el paso siguiente se determina una función de bondad de ajuste para evaluar cuánto se aproximan las distancias obtenidas a partir de *X* a las disparidades obtenidas de la transformación de esas distancias. Esta función se conoce con el nombre de *Stress* y tiene distintas definiciones. XLSTAT recoge varias expresiones:

Stress Normalizado

$$S = \sqrt{\frac{\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2}{\sum_i \sum_j \hat{d}_{ij}^2}}$$

Stress de Kruskal 1

$$S = \sqrt{\frac{\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2}{\sum_i \sum_j d_{ij}^2}}$$

Stress de Kruskal 2

$$S=\sqrt{\frac{\sum\sum_{i,j}\left(d_{ij}-\hat{d}_{ij}\right)^2}{\sum\sum_{i,j}\left(d_{ij}-\bar{d}\right)^2}}$$

Donde $\bar{d}$ es la media de las distancias representadas.

Para mayores valores de *Stress*, mejor será el ajuste encontrado entre distancias y disparidades. Es decir, el *Stress* no es propiamente un índice de bondad de ajuste, sino de «maldad» de ajuste. Su valor mínimo se encontrará, por tanto, en 0, cuando no exista diferencia entre distancias y disparidades. Su valor máximo no es estable, pero se conoce que su límite superior, para un número *n* de estímulos es:

$$\sqrt{1-\left(\frac{2}{n}\right)}$$

Como partimos de una matriz de coordenadas aleatoria, es de suponer que el ajuste nunca es muy bueno al principio. Por ello, se hace necesario llevar a cabo un proceso iterativo que vaya minimizando el valor del Stress.

Esto se consigue alterando los valores de las coordenadas de la matriz *X* de modo que la diferencia entre las distancias y disparidades derivadas a partir de ellos sea más pequeña ahora que en el paso anterior. La forma de llevar esto a cabo es sumar a la matriz *X* inicial una matriz de valores añadidos. Cada elemento de esta matriz contiene un valor que se sumará a la coordenada del estímulo *i* en la dimensión *a*. Este valor se determina mediante la expresión:

$$-a\left(\frac{\partial S}{\partial x_{ia}}\right)$$

a = constante que representa el tamaño del paso

$\left(\frac{\partial S}{\partial x_{ia}}\right)$ = derivada del Stress con respecto a la coordenada a-ésima del estímulo *i*

En el algoritmo de convergencia del proceso iterativo se utiliza otra función de *Stress*, conocida como *S*-Stress, cuya expresión es:

$$S-Stress=\sqrt{\frac{\sum_i\sum_j\left(d_{ij}-\hat{d}_{ij}\right)}{\sum_i\sum_j\hat{d}_{ij}^2}}$$

El valor de *Stress* es más alto cuanto mayor sea el número de estímulos, debido a que cuando tenemos pocos estímulos, el número de proximidades a ajustar en la solución será también pequeño, pero a medida que aumenta el número de estímulos, el número de proximidades a ajustar se incrementa rápidamente. El valor de *Stress* es siempre más alto para soluciones de menor dimensionalidad, e irá bajando a medida que la solución contenga un mayor número de dimensiones. Cuando el número de dimensiones es igual al número de estímulos menos 2(n-2), el ajuste será siempre perfecto. El objetivo en este caso será buscar un valor suficientemente bajo de *Stress* (buen ajuste) unido a una dimensionalidad también baja (representación parsimoniosa de los datos).

Alternativamente a *Stress* existe el índice RSQ para el ajuste del modelo a nuestros datos. Este índice es una correlación cuadrática entre las disparidades derivadas a partir de los datos originales, y las distancias derivadas por el modelo de escalamiento, de modo que puede ser interpretado como la proporción de varianza en las disparidades que es explicada por las distancias.

Su expresión es:

$$RSQ=\frac{\left[\sum_i\sum_j\left(d_{ij}-d_{\cdot\cdot}\right)^2\left(\hat{d}_{ij}-\hat{d}_{\cdot\cdot}\right)\right]^2}{\left[\sum_i\sum_j\left(d_{ij}-d_{\cdot\cdot}\right)^2\right]\left[\sum_i\sum_j\left(\hat{d}_{ij}-\hat{d}_{\cdot\cdot}\right)^2\right]}$$

Dado que su interpretación es mucho más sencilla y directa que la del *Stress*, y que sus límites son fijos (mínimo de cero y máximo de uno), Takane, Young y De Leew recomiendan apoyarse en este índice para la interpretación del ajuste de las soluciones proporcionadas.

En cuanto a los modelos de escalamiento para reducir la dimensión XLSTAT utiliza un algoritmo conocido como SMACOF («Scaling by majorizing a complicated function»). Su descripción excede a los objetivos de

este manual aunque el lector interesado puede consultarlo a través de la web: http://en.wikipedia.org/wiki/Stress_majorization

A continuación describiremos otros algoritmos habituales en otros paquetes estadísticos:

14.5. MODELO DE ESCALAMIENTO EN DIFERENCIAS INDIVIDUALES INDSCAL

INDSCAL supone una generalización del modelo euclídeo, de tal modo que obtiene una representación a partir de varias matrices de proximidades asumiendo que éstas difieren entre sí de forma sistemática y no aleatoria, tal y como supone un modelo replicado. Es decir, en lugar de considerar las diferencias entre matrices como sesgos en las respuestas de los sujetos, INDSCAL las contempla como diferencias porcentuales y cognitivas en el proceso de generación de las respuestas. Este modelo utiliza como entrada varias matrices de proximidades, por lo general, una por sujeto.

Cada proximidad $\hat{\delta}_{ijk}$ nos indicará la proximidad entre los estímulos *i* y *j* estimada por el sujeto *k*. Existen otras posibilidades, en las que las proximidades de cada matriz corresponden a una ocasión diferente, o a proximidades estimadas en diferentes condiciones o en base a atributos diferentes de los estímulos. El modelo considera que la relación entre proximidades y distancias es lineal:

$d_{ijk} = \sqrt{\sum_{a=1}^{m} w_{ka}(x_{ia} - x_{ja})^2}$ = peso del sujeto k-ésimo en la dimensión a-ésima.

El modelo INDSCAL puede considerarse como aquel en el que las diferencias individuales entre los sujetos surgen de las diferencias en los pesos otorgados a cada una de las distintas dimensiones que componen la solución común.

En la fórmula de la distancia anterior, si todos los pesos w_{ka} son iguales, la configuración de distancias entre estímulos para cada sujeto será la del grupo total, es decir, la solución común a todos los sujetos. A la configuración de distancias común a todos los sujetos se la conoce como el «espacio del grupo», y suele diferir de la configuración propia de cada sujeto. Cuando representamos las distancias entre estímulos en función del peso que cada una de las dimensiones tiene para un individuo concreto, la configuración de estímulos se verá «encogida» en aquellas dimensiones que tienen menor peso para el individuo. A esta configuración de distancias propia de este individuo se la conoce como «espacio del sujeto». Así pues,

podemos resumir el modelo INDSCAL diciendo que representa las diferencias entre los juicios emitidos por los sujetos en términos de la importancia que cada uno de ellos otorga a cada una de las dimensiones que componen la solución, pero todas las dimensiones son comunes a todos los sujetos.

El procedimiento subyacente en el modelo no métrico parte, al igual que el modelo métrico, de las proximidades, que se convierten en distancias ($d_{ij,k}$) absolutas mediante una constante aditiva.

Las distancias calculadas para cada sujeto se convierten luego en productos escalares $b_{ij,k}$, tales que:

$$b_{ij,k} = \sum_{a=1}^{r} y_{ia,k} y_{ja,k} \quad \text{con } y_{ia,k} = \sqrt{w_{ka}}\, x_{ia} \Rightarrow b_{ij,k} = \sum_{a=1}^{r} w_{ka} x_{ia} x_{ja}$$

Esta ecuación puede considerarse como un caso particular del modelo CANDECOMP (CANonical DECOMPosotion) para la descomposición de tablas de *N*-vías (3 en el caso del modelo INDSCAL). El modelo descompone una tabla de 3 vías y 3 modelos en un conjunto de parámetros para cada vía, que se combinan de forma multiplicativa para cada dimensión *a*, y de forma aditiva para el total de dimensiones. En el caso de INDSCAL, la segunda y tercera vías (representadas por los parámetros x_{ia} y x_{ja}) han de ser idénticas, pues se refieren al mismo conjunto de estímulos.

El uso del modelo CANDECOMP permite la estimación de los valores de los productos escalares $b_{ij,k}$ mediante regresión lineal, utilizando un algoritmo especial, el algoritmo de mínimos cuadrados alternantes (ALS, *Alternating Least Squares*). El algoritmo procede a estimar los valores de los parámetros w_{ka}, x_{ia} y x_{ja} por mínimos cuadrados, manteniendo uno de ellos fijo y los otros dos libres, de forma alternante.

Cuando, transcurridas una serie de iteraciones, el ajuste entre los datos y la solución es satisfactorio, se fija el mismo valor para la segunda y tercera vías y se estima el valor de la primera vía (representada por el parámetro w_{ka}).

La salida ofrecida por el modelo presenta una primera matriz de coordenadas $X(n \times r)$, semejante a las de los modelos métrico y no métrico. Esta matriz representa el espacio de los estímulos para el total de los sujetos (espacio del grupo). La salida también ofrece una segunda matriz de pesos $W(m \times r)$, que contiene los pesos otorgados por cada uno de los m sujetos a cada una de las r dimensiones.

Esta matriz representa el espacio de los sujetos. La denominación de espacio de estímulos y espacio de sujetos debe tomarse en el sentido de que son dos espacios distintos, por lo que no es posible representar ambos en un único gráfico. El espacio de sujetos tiene una serie de propiedades interesantes para la interpretación de la solución proporcionada por INDSCAL.

Por ejemplo, se cumple que si elevamos al cuadrado el peso otorgado por un sujeto a una dimensión determinada, el valor obtenido se corresponde con la proporción de varianza en los datos del sujeto que es explicada por esa dimensión.

También se cumple que si sumamos todos los pesos al cuadrado para un mismo sujeto, el valor obtenido es la proporción de varianza en los datos del sujeto que es explicada por la solución proporcionada por INDSCAL, es decir, este valor coincide con el del estadístico RSQ para ese sujeto.

Por otra parte, dado que sólo se permiten pesos positivos, la presencia de valores negativos en la matriz W puede indicar un mal ajuste del modelo a los datos. No obstante, si los valores son muy pequeños pueden tomarse simplemente como aproximaciones a un valor cero en el peso. En este último caso, no existe ningún problema de ajuste.

Adicionalmente, A partir de las dos matrices anteriores (X y W) es posible recuperar el espacio de estímulos individual para cada uno de los sujetos (espacio del sujeto). Esto se consigue simplemente multiplicando cada coordenada del espacio de estímulos total por la raíz cuadrada del peso asignado por el sujeto a esa dimensión (ver la penúltima fórmula mostrada). Las nuevas coordenadas, $y_{ia,k}$, muestran el espacio del grupo «encogido» en aquellas dimensiones que resultan ser menos relevantes para el individuo.

14.6. MODELO DE ESCALAMIENTO DESDOBLADO UNFOLDING

Existen también modelos de MDS para matrices de datos que no son cuadradas. Todos los modelos de escalamiento que hemos visto hasta ahora trabajaban con matrices cuadradas, con el mismo número de estímulos en filas y columnas. También hemos utilizado matrices cuadradas al aplicar el modelo INDSCAL, aunque las tratamos como una única matriz en tres vías y dos modos. Las matrices cuadradas (especialmente si son simétricas) constituyen el tipo de datos de entrada más común en MDS.

Alternativamente, la característica fundamental de una matriz rectangular es que las entidades representadas en las filas y las columnas (generalmente sujetos y estímulos) son diferentes. Por tanto, un análisis MDS con una matriz rectangular deberá representar conjuntamente ambas enti-

dades. Esto representa una propiedad sumamente interesante de este tipo de modelos. Recordemos que INDSCAL representa un espacio para los sujetos y otro para los estímulos, pero no ambos conjuntamente.

En los modelos de escalamiento con matriz rectangular los datos de entrada suelen ser puntuaciones de preferencia otorgados por un grupo de sujetos para un conjunto de estímulos, aunque también es posible utilizar otro tipo de puntuaciones.

Suelen utilizarse dos tipos de modelos para matrices rectangulares de preferencia: el modelo vectorial y el modelo del «punto ideal» (que aquí denominamos «desdoblado» o *unfolding*). En el primero, una de las entidades (generalmente los estímulos) se representa como puntos en un espacio, mientras que la otra (generalmente los sujetos) se representa como vectores en ese mismo espacio.

De este modo, la proyección de las posiciones de los estímulos sobre el vector de un sujeto (cuyo extremo indica la máxima preferencia) deberá reflejar las preferencias de ese sujeto. En el modelo desdoblado, tanto los sujetos como los estímulos se representan como puntos. Los puntos que representan a los sujetos en el espacio de la solución indican la zona donde reencontraría la máxima preferencia de cada sujeto, de tal modo que a medida que nos alejamos de uno de esos puntos en cualquier dirección, la preferencia va disminuyendo. Si alguno de los estímulos está próximo a un punto, ese estímulo es el ideal para el sujeto representado por el punto. Por esta razón se conoce a este modelo como el modelo del «punto ideal».

La matriz de entrada en el MDS desdoblado es una matriz rectangular de preferencias en dos vías y dos modos (generalmente *sujetos* × *estímulos*), donde cada entrada de la matriz corresponde a la preferencia expresada por el sujeto *i* por el estímulo *j*. Dado que cada fila de la matriz contiene las puntuaciones de preferencia de una fuente de datos distinta (generalmente un sujeto), es habitual suponer aquí que los datos de cada fila son condicionales.

En el MDS desdoblado, la matriz rectangular es un trozo de la diagonal de una matriz de proximidades incompleta, lo que implica que el análisis da por perdida gran cantidad de la información contenida en la matriz de proximidades completa. Luego el modelo desdoblado es el modelo más propenso a no converger o proporcionarnos soluciones degeneradas. El modelo asume que la proximidad del estímulo *j* al punto ideal del sujeto *i* (p_{ij}) es una fundición de la preferencia del sujeto i-ésimo por el estímulo j-ésimo (p_{ij}). Tenemos:

$$\pi_{ij} = f(p_{ij}) = d_{ij}^2 \text{ con } d_{ij}^2 = \sum_{s=1}^{r} (y_{is} - x_{js})^2$$

y_{ia} = coordenada del sujeto i-ésimo en la dimensión a-ésima.

x_{ja} = coordenada del estímulo j-ésimo en la dimensión a-ésima.

f puede ser lineal (caso métrico) o monotónica (caso no-métrico).

14.7. MODELO DE ESCALAMIENTO CON REPLICACIÓN

En el modelo de escalamiento con replicación se trata la matriz de entrada, que es una matriz en tres vías, como varias replicaciones de una misma matriz en dos vías. El ajuste del modelo a las *m* matrices se calcula mediante una variante de *S-Stress* basada en la media de la razón entre las sumas de cuadrados del error y las sumas de cuadrados totales para cada matriz. Tenemos:

$$S-\text{Stress} = \sqrt{\frac{1}{m}\sum_{m}\frac{\sum_{i}\sum_{j}\left(d_{ij}-\hat{d}_{ij}\right)^2}{\sum_{i}\sum_{j}\hat{d}_{ij}^2}}$$

Para mostrar el ajuste final del modelo a los datos se utilizará *Stress* y la medida RSQ promedio para las *m* matrices de datos. Adicionalmente se mostrarán también los valores de ajuste para cada matriz individual.

14.8. MODELOS GEMSCAL E IDIOSCAL

El modelo GEMSCAL (*Generalizad Euclidean Model SCALing*) propuesto por Young puede ser asimilado al más conocido modelo IDIOSCAL (*Individual Differences in Orientation SCALing*) propuesto por Carrol y Chang. El modelo expresa la existencia de diferencias entre las fuentes de datos (generalmente sujetos) permitiendo que cada fuente lleve a cabo una rotación diferente de las dimensiones del espacio de estímulos común.

Esta es la principal diferencia entre el modelo GEMSCAL y el modelo INDSCAL, donde la orientación del espacio de estímulos es única. Podemos considerar, pues, a INDSCAL como un caso particular del modelo GEMSCAL.

El modelo utiliza como entrada generalmente varias matrices de proximidades, aunque en versiones más complejas del mismo también pueden utilizarse matrices rectangulares y matrices asimétricas. La familia GEMSCAL contiene en realidad 40 modelos diferentes, 20 de los cuales son para matrices cuadradas (4 para matrices simétricas y 16 para matrices asimétricas) y

otros 20 son para matrices rectangulares. Dada esta complejidad nos centraremos en el caso de varias matrices de proximidades simétricas como entrada, que es de hecho el modelo IDIOSCAL.

En el modelo IDIOSCAL, la distancia entre dos estímulos *i* y *j* para la matriz *k* (la k-ésima fuente de datos) viene dada por la siguiente expresión:

$$d_{ijk} = \sqrt{\sum_{a=1}^{m}\sum_{a'=1}^{m'}(x_{ia} - x_{ja})w_{kaa'}(x_{ia'} - x_{ja'})}$$

Los subíndices *a* y *a'* representan las *m* dimensiones correspondientes, respectivamente, al espacio común de estímulos y al espacio de cada sujeto (o fuente de datos). La matriz W_{kaa}, es una matriz de dimensiones $m \times m$ positiva definida o semidefinida, que contiene los pesos asociados con cada una de las *k* matrices correspondientes a las distintas fuentes de datos. A efectos prácticos, lo que proporciona esta matriz de pesos es una rotación ortogonal del espacio de estímulos a un nuevo sistema de coordenadas específico de cada fuente de datos. Si consideramos el caso especial donde la matriz W_{kaa}, es una matriz diagonal con pesos no negativos, entonces el modelo GEMSCAL pasa a simplificarse y convertirse en el modelo INDSCAL. En efecto, si w_{kaa}, = w_{ka}, entonces $a = a'$, por lo que el producto $(x_{ia} - x_{ja})(x_{ia'} - x_{ja'})$ se convierte en $(x_{ia} - x_{ja})^2$, y la fórmula de la distancia pasa a ser la ya conocida del modelo INSCAL:

$$d_{ijk} = \sqrt{\sum_{a=1}^{m} w_{ka}(x_{ia} - x_{ja})^2}$$

La representación final es un espacio conjunto para estímulos y sujetos, donde los estímulos aparecen representados como puntos, y los sujetos como vectores. Las direcciones a las que apuntan los vectores de un sujeto en el espacio corresponden a las direcciones más importantes para ese sujeto, mientras que la longitud de los vectores corresponderá a la importancia que ese sujeto otorga cada dirección. La proyección de los estímulos sobre los vectores de un sujeto proporcionará el espacio de estímulos propio de ese sujeto, donde cada dimensión se verá «encogida» en función de la longitud del vector correspondiente.

14.9. MODELOS PARA MATRICES ASIMÉTRICAS

Es habitual en el MDS que la matriz cuadrada de proximidades sea simétrica $(\delta_{ij} = \delta_{ji})$. Pero en la práctica nos podemos encontrar con la po-

sibilidad de que existan asimetrías en las proximidades (por ejemplo, una situación de interacciones sociales donde el sujeto A puede dirigirse al B más a menudo de lo que el sujeto B se dirige al A). En ese tipo de situaciones es posible analizar por separado cada mitad triangular de la matriz de proximidades (obtendríamos una solución para las interacciones en un sentido y otra solución para las interacciones en sentido inverso) o promediar los resultados para ambas matrices triangulares y utilizar la matriz promedio como entrada para un análisis MDS (la proximidad entre los sujetos A y B será ahora el promedio de interacciones en ambos sentidos) o incluso utilizar ambas matrices triangulares como entrada y tratarlas como replicaciones (la solución mostrará una solución común a ambas, así como el grado de acuerdo entre las presentaciones derivadas de ambas matrices). No obstante, existen modelos de MDS apropiados para trabajar con datos asimétricos. Los más importantes y utilizados son: el modelo ASCAL (*Asymmetric SCALing*) para datos en dos vías y un modo, y el modelo AINDS (*Asymmetric Individual Differences Scaling*), para datos en tres vías y dos modos.

Modelo ASCAL

Este modelo toma como entrada una matriz de proximidades asimétrica en dos vías y un modo. La distancia entre los estímulos *i* y *j* viene dada por:

$$d_{ij} = \sqrt{\sum_{a=1}^{m} v_{ia}\left(x_{ia} - x_{ja}\right)^2}$$

v_{ia} = matriz de pesos de dimensiones $n \times m$

Las celdillas de v_{ia} indican el peso de cada uno de los *n* estímulos en cada una de las *m* dimensiones. La salida del modelo ASCAL contendrá una matriz *X* de coordenadas de los *n* estímulos en las *m* dimensiones y una matriz *V* de pesos de los *n* estímulos en las *m* dimensiones.

Modelo AINDS

Este modelo toma como entrada una matriz de proximidades asimétrica en tres vías y dos modos. Se asume que las distancia entre los estímulos *i* y *j* para la matriz *k* (k-ésima fuente de datos) viene dada por la siguiente expresión:

$$d_{ijk} = \sqrt{\sum_{a=1}^{m} v_{ka} w_{ka}\left(x_{ia} - x_{ja}\right)^2}$$

v_{ia} es una matriz de pesos, de dimensiones $n \times m$ cuyas celdas indican el peso de cada uno de los n estímulos en cada una de las m dimensiones.

w_{ka} es una matriz de dimensiones $r \times m$ cuyas celdas indican el peso otorgado por cada sujeto (o fuente de datos) a cada dimensión. Esta matriz tiene una interpretación similar a la matriz correspondiente del modelo INDSCAL.

La salida del procedimiento AINDS presenta una matriz X de coordenadas de los n estímulos en las m dimensiones, una matriz V de pesos de los n estímulos en las m dimensiones y una matriz W de pesos de los r sujetos en las m dimensiones.

EJERCICIO 14-1. El fichero MARCAS_SOLIDARIAS contiene información referente a los valores medios que los trabajadores encuestados en estas empresas dieron a diferentes valores y características de la empresa en la que trabajan.

Se pide realizar un análisis de escalamiento multidimensional con el objetivo de posicionar nuestra marca X respecto a las demás.

Lo primero que necesitamos calcular es la matriz de similitud/disimilitud entre los datos. Para ello utilizaremos XLSTAT → Descripción de Datos → Matriz de similitud / disimilitud (Figura 14.1) poniendo especial atención en calcular las proximidades para las filas y en pedir en la pestaña de resultados la matriz de proximidad.

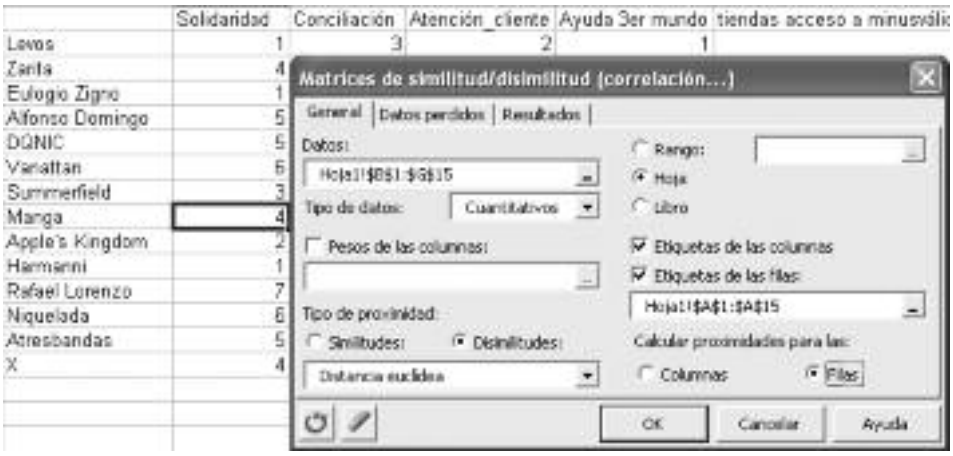


FIGURA 14-1

Los resultados obtenidos permiten abordar ya el escalamiento multidimensional.

	Levos	*Zarita*	*Eulogio Zigno*	*Alfonso Domingo*	*DQNIC*	*Vanattan*		*X*
Levos	**0,0**	8,9	7,5	10,1	5,0	5,7	…	6,4
Zarita	8,9	**0,0**	10,6	5,6	6,4	8,1	…	4,6
Eulogio Zigno	7,5	10,6	**0,0**	8,2	8,7	8,7	…	6,6
Alfonso Domingo	10,1	5,6	8,2	**0,0**	7,9	8,1	…	4,7
DQNIC	5,0	6,4	8,7	7,9	**0,0**	3,6	…	4,0
Vanattan	5,7	8,1	8,7	8,1	3,6	**0,0**	…	5,6
…	…	…	…	…	…	…	…	
X	6,4	4,6	6,6	4,7	4,0	5,6	…	**0,0**

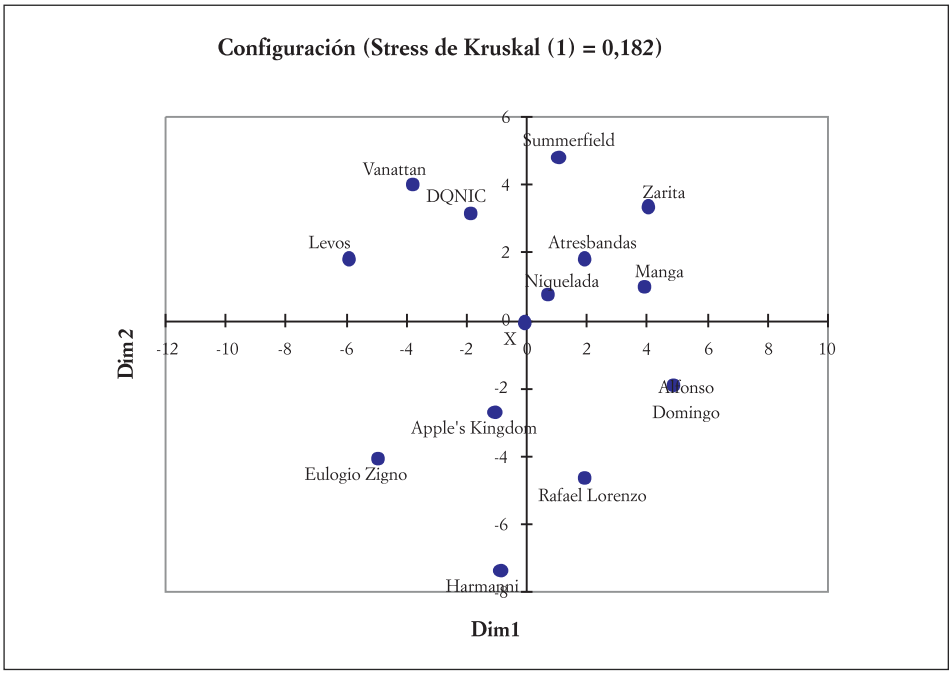
Para llevar a cabo el escalado multidimensional seleccionamos XLSTAT → Análisis de Datos → Multidimensional Scaling (Figura 14.2).

FIGURA 14-2

Los resultados del análisis son los siguientes:

Resultados para un espacio de dimensión de 2 dimensiones

	Dim 1	*Dim 2*
Levos	-5,875	1,852
Zarita	4,015	3,325
Eulogio Zigno	-4,964	-4,075
Alfonso Domingo	4,922	-1,882
DQNIC	-1,847	3,179
Vanattan	-3,770	4,028
Summerfield	1,042	4,816
Manga	3,928	0,972
Apple's Kingdom	-1,051	-2,727
Harmanni	-0,836	-7,363
Rafael Lorenzo	1,888	-4,632
Niquelada	0,723	0,761
Atresbandas	1,889	1,800
X	-0,064	-0,052



MAPA PERCEPTUAL

Así, es fácil comprobar como si bien las marcas en cada cuadrante son más parecidas entre sí, nuestra marca aparece centrada por lo que en función de las decisiones que tomemos será fácil evolucionar hacia uno u otro lado del cuadrante.

BIBLIOGRAFÍA

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. y Tatham, R. L. (2006). Multivariate Data Analysis. Sixth Edition. Pearson, Prentice Hall. New Jersey.

Pérez López, C. (2005). Métodos estadísticos avanzados con SPSS. Thomson Paraninfo.

Pérez López, C. (2005). Técnicas estadísticas con SPSS12. Aplicaciones al análisis de datos. Pearson Alambra.

Pérez López, C. (2004). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS. Pearson Alhambra.

CAPÍTULO XV
MODELO DE REGRESIÓN LINEAL MÚLTIPLE

JORGE ONRUBIA FERNÁNDEZ

15.1. INTRODUCCIÓN

En sus diferentes áreas de especialización, la teoría económica recoge proposiciones dirigidas a explicar fenómenos con trascendencia económica o el origen de determinadas decisiones o comportamientos con trasfondo económico, Por regla general, estas proposiciones encierran *relaciones determinísticas* entre variables. De este modo, el resultado buscado a través del valor de la variable explicada viene determinado con completa certidumbre por la estructura funcional que incorpora las variables explicativas elegidas en el modelo propuesto. Como ejemplos podemos citar la ecuación de oferta de trabajo, la ecuación de demanda de un determinado bien de consumo o la función de producción de cualquier bien o servicio. Asimismo, estas relaciones se extienden al análisis de programas públicos. La conocida ecuación de Mincer que explica el salario a partir del nivel de educación y la experiencia laboral o el modelo que estudia la influencia de las ayudas para comedores escolares sobre la oferta de trabajo de las mujeres casadas constituyen ejemplos en este ámbito.

Para determinar la validez de estas proposiciones teóricas, el análisis econométrico tiene por objeto proporcionar evidencia empírica a partir de los datos disponibles correspondientes a las variables que intervienen en el modelo. Sin embargo, a diferencia de lo que sucede en el modelo teórico, *la naturaleza del modelo econométrico es estocástica*. Dos razones están detrás de este hecho. Por un lado, los modelos teóricos incorporan aquellas variables explicativas consideradas esenciales, pero en el mundo real –del que no olvidemos proceden los datos empleados en el análisis empírico– existen otros factores que condicionan la influencia de estas variables y que no es factible identificar de antemano. Por otro lado, también es frecuente encontrar problemas de medición de las variables incluidas en los modelos teóricos, pudiéndose hablar en algunos casos incluso de imposibilidad de contar con mediciones reales, como sucede con el valor de los servicios obtenidos del uso de bienes de consumo duradero o la renta de ciclo vital. Ante estas limitaciones, la especificación econométrica del modelo ha de incluir un factor estocástico cuya misión es incorporar al modelo teórico la aleatoriedad presente en el mundo real, si bien como vere-

mos más adelante *esta aleatoriedad no puede enmascarar cualquier error de especificación del modelo*.

De acuerdo con lo expuesto, el punto de partida para la realización un análisis econométrico debe ser la correcta definición del hecho o actuación que se desea estudiar. Esto supone la propuesta de un modelo teórico explicativo, con una especificación concreta de la relación existente entre las variables que se consideran relevantes para explicar la cuestión de estudio. En esta primera fase, los postulados teóricos que nutren las distintas áreas de estudio de la teoría económica y de la economía aplicada deben servir para dotar de consistencia a los modelos propuestos.

El trabajo econométrico debe iniciarse con la búsqueda de la mejor información estadística disponible para conseguir una medición precisa de las variables que hemos decidido han de intervenir en el modelo explicativo. No hay que olvidar, como se ha dicho, que las limitaciones existentes para obtener información sobre algunas variables obligan en ocasiones a utilizar otras *variables sustitutivas o aproximadas* con la intención de capturar los elementos esenciales de las originales (*variables proxy*). En relación con los datos empleados aparecen varios tipos de problemas que afectan tanto a la estimación econométrica de los modelos como a la interpretación de los resultados de las regresiones. El más habitual es la *multicolinealidad*, y aparece ante la correlación entre variables explicativas, lo que imposibilita identificar con precisión la influencia de cada una de ellas en la variable explicada. Otros problemas tienen su origen en la *ausencia de datos* en las muestras empleadas o en la sustitución de la información individual de cada dato por *valores derivados de la agregación de observaciones* (la práctica más común es la utilización del valor medio de una serie). La existencia de observaciones con valores extremos en las muestras también puede introducir imperfecciones en las estimaciones. Como veremos en este capítulo, la econometría ofrece métodos diversos para paliar estos problemas.

A continuación, el siguiente paso es la especificación del modelo econométrico a estimar. Obviamente, éste debe tratar de recoger de la forma más fiel posible el modelo teórico propuesto para explicar el fenómeno económico que ocupa nuestra atención investigadora. No obstante, la elección, por ejemplo, de una relación lineal entre las K variables explicativas frente a una alternativa no lineal condiciona el modelo econométrico a utilizar, con la consecuente elección de la metodología de estimación. Asimismo, la disposición de los datos, referidos a un determinado periodo de tiempo (*datos de sección cruzada o cross section*), estructurados longitudinalmente para periodos de tiempo consecutivos (*datos en series temporales*), o en una combinación de ambas presentaciones en forma de *datos de panel*, igualmente condiciona los métodos de estimación a utilizar.

El proceso de estimación del modelo econométrico tiene como resultado principal la obtención de los correspondientes parámetros regresores –las incógnitas del problema de regresión– que definen la forma funcional precisa del modelo formulado de acuerdo con la información suministrada por los datos disponibles. El cómputo de estos parámetros depende del método de estimación elegido (mínimos cuadrados ordinarios, máxima verosimilitud, mínimos cuadrados generalizados, etc.), lo que afectará a su valor numérico. La exigencia de dotar al modelo econométrico del necesario componente estocástico también obliga a garantizar la estricta aleatoriedad del término de error, lo que como veremos puede suponer un proceso de revisión del modelo inicial a estimar, por ejemplo, al identificarse errores de especificación derivados de la posible omisión de variables explicativas relevantes o de la introducción de alguna irrelevante.

El análisis concluye con la selección de las estimaciones econométricas realizadas. De hecho, el análisis econométrico además de buscar la forma precisa de relación entre las variables que definen el modelo propuesto cumple una misión de validación empírica de éste. La obtención de distintos contrastes de significación estadística referidos a las variables cuyos regresores se estiman y a la propia consistencia del ajuste funcional que supone la ecuación de regresión permite validar tanto la idoneidad de las variables seleccionadas como la correcta especificación del modelo. Se abre así un proceso de interacción entre el análisis teórico y el empírico que debe culminar con la aceptación o rechazo de la capacidad explicativa del modelo propuesto.

15.2. EL MODELO DE REGRESIÓN LINEAL GENERAL

El *modelo de regresión lineal clásico o general* es el más utilizado y puede considerarse como el modelo básico para el estudio de la mayor parte de los métodos de estimación de ecuaciones econométricas. Habitualmente suele diferenciarse entre el *modelo de regresión lineal simple* y *múltiple*. La diferencia entre ambos se limita a la consideración, en el primer caso, de una relación lineal con una única variable explicativa, además de la participación de un término constante, mientras que en el segundo, se incorporarían más variables explicativas.

El modelo de regresión lineal simple

El *modelo de regresión lineal simple* plantea una relación lineal entre una variable dependiente o explicada y y una variable independiente

o explicativa *x*. La forma funcional genérica del modelo de regresión lineal simple es:

$$y_i = \alpha + \beta x_i + \varepsilon \qquad [1]$$

El subíndice *i* identifica la *i*-ésima observación de las *N* observaciones muestrales de cada variable. Podemos ver que la condición de linealidad impuesta a la relación funcional define la ecuación econométrica [1] como una recta en la que el parámetro *a*, (término independiente) establece la intersección con el eje de ordenadas, mientras que el parámetro *b* se identifica con su pendiente y, por tanto, determina la incidencia de la variable exógena *x* en la variable endógena *y*. El término ε es la perturbación aleatoria que impone el carácter estocástico del modelo econométrico. Esta perturbación aleatoria recoge aditivamente el término de error del modelo esperable en la medida que los datos reales empleados en la estimación no es previsible que respondan a una relación determinística estable. Como hemos adelantado, errores de medida en las variables o la imposibilidad de identificar otros factores no esenciales que influyen en los valores de *y* explican la razón de ser de esta perturbación aleatoria y, por ende, el carácter estocástico del modelo de regresión.

La notación matricial de la ecuación [1] nos ayudará a entender la incorporación de los datos empíricos al análisis de regresión:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad [2]$$

Comencemos por ordenar los *N* pares de observaciones (y_i, x_i) que integran la muestra de datos correspondientes a las variables que intervienen en el modelo. Así, las observaciones de la variable explicada *y* se recogen en un vector columna de dimensión $N \times 1$, mientras que para la variable exógena *x* y el término constante *a*, establecemos una matriz de orden $N \times 2$:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_{N-1} \\ y_N \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \dots \\ 1 & x_{N-1} \\ 1 & x_N \end{pmatrix} \qquad [3]$$

La matriz **X** recoge el valor 1 en todos los elementos de su primera columna, lo que equivale a considerar el término constante *a* como una variable explicativa más. Para ilustrar esta formulación emplearemos la siguiente muestra (Tabla 1.a) integrada por una serie de 20 observaciones de

las variables «gasto anual en alimentación y vestido» (GAV) y «renta disponible del hogar» (RDH), ambas expresadas en euros y correspondientes a hogares españoles en el año 2005. En la Tabla 1.b se ofrece su presentación alternativa de acuerdo con la expresión [3].

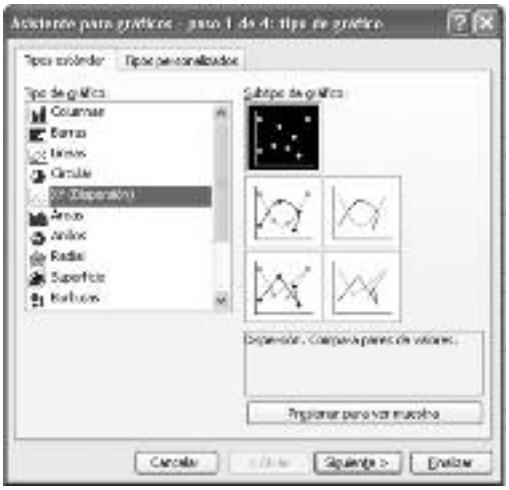
	A	B	C	D
1	obs.	GAV	RDH	
2	1	9908	20022	
3	2	8084	14864	
4	3	5862	12009	
5	4	10153	21139	
6	5	9696	20012	
7	6	10640	19326	
8	7	5673	11503	
9	8	8953	16293	
10	9	9803	17563	
11	10	9449	15418	
12	11	5359	9566	
13	12	8606	13820	
14	13	4820	8263	
15	14	11104	18499	
16	15	9258	14112	
17	16	7831	15050	
18	17	7194	13667	
19	18	7735	12796	
20	19	9196	14140	
21	20	7696	11586	
22				

TABLA 1.A

	A	B	C	D	E
1	GAV		a i	RDH	
2	9908		1	20022	
3	8084		1	14854	
4	5862		1	12009	
5	10153		1	21139	
6	9696		1	20012	
7	10640		1	19326	
8	5673		1	11503	
9	8953		1	16293	
10	9803		1	17563	
11	9449		1	15418	
12	5359		1	9566	
13	8606		1	13820	
14	4820		1	8263	
15	11104		1	18499	
16	9258		1	14112	
17	7831		1	15050	
18	7194		1	13667	
19	7735		1	12796	
20	9196		1	14140	
21	7696		1	11586	
22					

TABLA 1.B

Empleando el *Asistente para Gráficos* de Excel podemos observar la relación existente entre ambas variables. Para ello eligimos el Tipo de Gráfico «XY (Dispersión) y el Subtipo de Gráfico «Dispersión. Compara pares de valores». Así, en el paso 4 de 4,



Obtenemos el Gráfico 1 que muestra la relación existente entre las variables "x" (RDH) e "y" (GAV) para cada una de las 20 observaciones de la muestra:

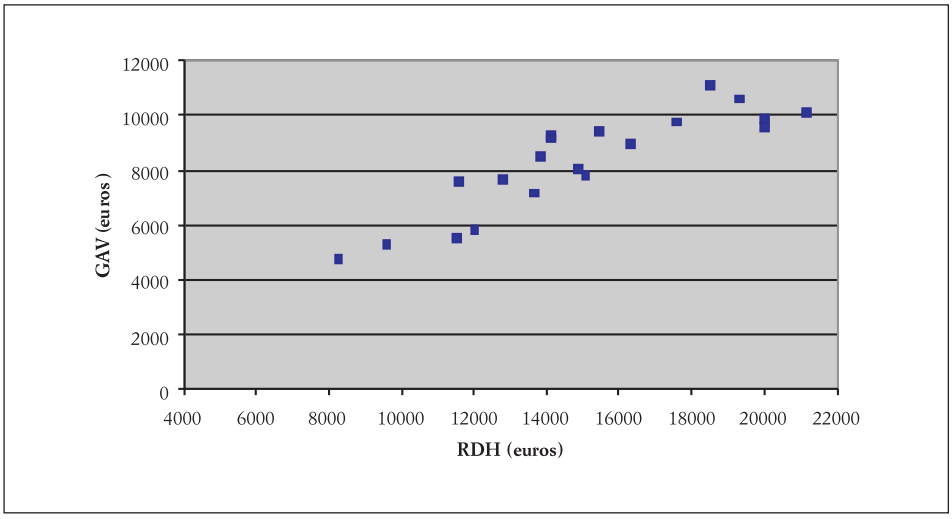


FIGURA 1. *Gasto en Alimentación y Vestido - Renta Disponible del Hogar (2005)*

Con los datos de esta muestra, podemos tratar de encontrar una relación funcional simple entre la renta disponible de los hogares (RDH) y su gasto correspondiente en alimentación y vestido (GAV). De acuerdo con los postulados de la teoría económica el gasto de consumo en bienes básicos como la alimentación y el vestido de los hogares podría explicarse fundamentalmente por el nivel de renta disponible de esos hogares. Inicialmente, podemos presumir que esta relación es lineal, tal que,

$$GAV = \alpha + \beta \cdot RDH \qquad [4]$$

donde el parámetro a representa el nivel de consumo autónomo (o consumo de supervivencia para un nivel de renta nulo y, por tanto, $a > 0$), mientras que b recogería la propensión media a consumir del hogar (lo que determina que su valor esté comprendido entre 0 y 1). En consecuencia, la relación teórica vendría definida por una recta con pendiente positiva, proposición que de acuerdo con la información que nos suministran los datos representados en el gráfico 1 parece, al menos inicialmente, aceptable.

Sin embargo, hemos de aceptar que la función lineal formulada en [4] es solamente una aproximación. Como ya se expuso, incluso aceptando que la renta disponible influyese de forma capital en el montante de gasto dedi-

cado por los hogares al consumo de alimentos y vestido, es difícil imaginar una relación determinística entre ambas variables. De hecho, en el mundo real se producen influencias en las decisiones de consumo no predecibles y, por tanto, no incorporables en el modelo teórico propuesto. Se trata, por tanto, de perturbaciones aleatorias que influyen en la variable *GAV* y que no deben ser interpretados como la consecuencia de no haber incorporado al modelo todos los factores explicativos. En la medida que, dada su naturaleza estocástica, esos *shocks* aleatorios no pueden ser anticipados, su impacto sobre la variable explicada no puede establecerse por la relación de las variables explicativas. Como veremos más adelante, la ausencia de variables explicativas relevantes simplemente nos situaría ante una incorrecta especificación del modelo, nunca ante el origen del componente aleatorio del modelo.

¿Cómo podemos incorporar al modelo planteado esa aleatoriedad propia del mundo real? La forma más usual es suponer que, de acuerdo con la expresión [1], el modelo propuesto tiene dos componentes, uno determinístico (el recogido en [4]) y otro aleatorio e, y cuya relación es aditiva. Así, el modelo propuesto en [4] puede redefinirse en términos estocásticos como:

$$GAV = \alpha + \beta \cdot RDH + e \qquad [5]$$

Esta especificación estocástica del modelo nos permite ahora explicar los niveles de *GAV* a partir de los correspondientes niveles de la *RDH* y de shocks aleatorios no determinables *a priori*. Precisamente, el objetivo del análisis econométrico es ofrecer una estimación empírica de los parámetros a y b , incógnitas del modelo planteado.

Regresión por Mínimos Cuadrados Ordinarios (MCO)

El procedimiento más empleado en el análisis econométrico para la estimación de los parámetros del modelo de regresión lineal es el *método de Mínimos Cuadrados Ordinarios* (en adelante MCO). En el modelo lineal simple (con una única variable explicativa), el método consiste en ajustar una recta a la nube de puntos formada por todos los pares de observaciones y_i, x_i de la muestra (gráfico 1) con la condición de que la suma de los cuadrados de las distancias entre los valores de y que proporciona la recta estimada $(\hat{y}_i)$ y los verdaderos de los datos observados (y_i) sea mínima. La consideración de las distancias al cuadrado evita el problema de los valores negativos que surgen en las estimaciones por exceso de la variable explicado $(\hat{y}_i > y_i)$. Los postulados econométricos demuestran la utilización del método de MCO como criterio de ajuste proporciona estimadores con propiedades estadísticas satisfactorias.

Hay que tener en cuenta, no obstante, que lo que se obtiene del procedimiento de regresión son estimaciones de esos parámetros, es decir, $\hat{\alpha}$ para α, $\hat{\beta}$ para β, y para e. Por consiguiente, podemos distinguir entre la regresión poblacional, definida como $E[y_i|x_i] = \alpha + x_i \cdot \beta$, y su correspondiente estimación muestral, $\hat{y}_i = \hat{\alpha} + x_i \cdot \hat{\beta}$. Así, la perturbación aleatoria correspondiente a la *i*-ésima observación será $\varepsilon_i = y_i - (\alpha + x_i \cdot \beta)$, mientras que su estimador $\hat{\varepsilon}$ se obtiene a partir de las distancias (residuos) entre el valor observado para y_i y el estimado en el proceso de ajuste $\hat{y}_i$, tal que $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + x_i \cdot \hat{\beta})$.

El Gráfico 2 ilustra esta relación entre la regresión poblacional y la resultante del proceso de estimación. En el ejemplo propuesto, las incógnitas a resolver en el proceso de regresión son los parámetros *a*, *b* y *e* que definen la relación estocástica [5] entre GAV –la variable dependiente, *y*– y RDH –la variable explicativa, *x*–.

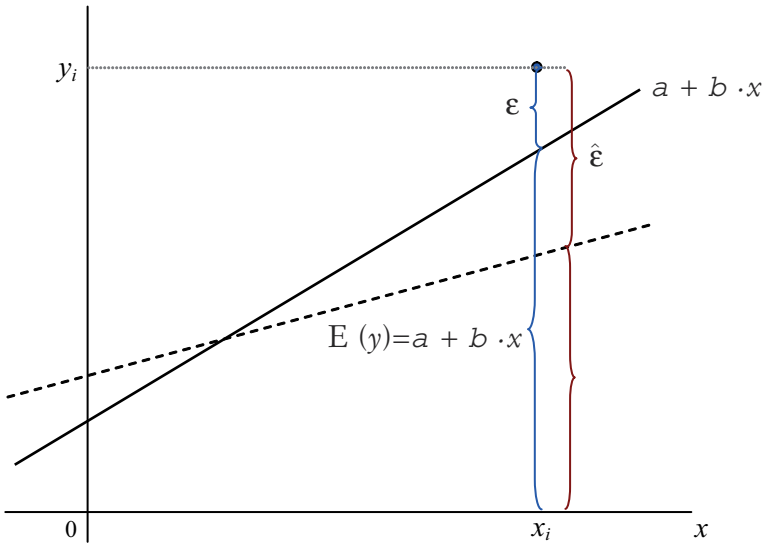


GRÁFICO 2

Como hemos señalado, el método de estimación MCO exige que los parámetros obtenidos hagan mínima la suma de los cuadrados de los resi-

duos. Por tanto, en el modelo lineal simple con una única variable explicativa, los parámetros estimados $\hat{\alpha}$ y $\hat{\beta}$ deben asegurar:

$$\min \sum_{i=1}^{N} \hat{\varepsilon}_i = \min \sum_{i=1}^{N} \left(y_i - \hat{\alpha} - \hat{\beta} \cdot x_i\right)^2 \qquad [6]$$

Para obtener los coeficientes MCO $\hat{\alpha}$ y $\hat{\beta}$ según este método de ajuste, en primer lugar deben obtenerse las condiciones de primer orden (CPO) correspondientes al problema de minimización formulado en [6]:

$$\text{CPO}(\hat{\alpha}): \frac{\partial\left[\sum_{i=1}^{N} \hat{\varepsilon}_i^2\right]}{\partial \hat{\alpha}} = -\sum_{i=1}^{N} 2\left(y_i - \hat{\alpha} - \hat{\beta} \cdot x_i\right) = 0 \mapsto \sum_{i=1}^{N} \hat{\varepsilon}_i = 0 \qquad [7]$$

$$\text{CPO}(\hat{\beta}): \frac{\partial\left[\sum_{i=1}^{N} \hat{\varepsilon}_i^2\right]}{\partial \hat{\beta}} = \sum_{i=1}^{N} 2\left(y_i - \hat{\alpha} - \hat{\beta} \cdot x_i\right) \cdot \left(-x_i\right) = 0 \mapsto \sum_{i=1}^{N} x_i \hat{\varepsilon}_i = 0 \qquad [8]$$

Operando en estas dos CPO se forma el siguiente sistema de ecuaciones en el que $\hat{\alpha}$ y $\hat{\beta}$ son las incógnitas buscadas:

$$\begin{aligned} \sum_{i=1}^{N} y_i &= N \cdot \hat{\alpha} + \hat{\beta} \cdot \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i &= \hat{\alpha} \cdot \sum_{i=1}^{N} x_i + \hat{\beta} \cdot \sum_{i=1}^{N} x_i^2 \end{aligned} \qquad [9]$$

Para solucionar el sistema y obtener el valor del término independiente de la recta dividimos ambos términos de la primera ecuación entre el número de observaciones *N*. Esto nos permite además comprobar que la recta de regresión ajustada por el método MCO pasa por los puntos correspondientes a los valores medios tanto de la variable dependiente como de la explicativa [1]:

$$\frac{\sum_{i=1}^{N} y_i}{N} = \hat{\alpha} + \hat{\beta} \cdot \frac{\sum_{i=1}^{N} x_i}{N} \qquad [10]$$

[1] Esto siempre se cumple salvo en el caso de que no exista intersección con el origen.

A partir de los valores medios de ambas variables $(\bar{x}, \bar{y})$, despejando en [10] expresamos directamente el valor del coeficiente $\hat{\alpha}$.

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \qquad [11]$$

Sustituyendo [11] en la segunda ecuación de [9] podemos calcular el valor del parámetro $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i - \hat{\alpha} \cdot \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2} = \frac{\left(\sum_{i=1}^{N} x_i y_i\right) - N\bar{x}\bar{y}}{\left(\sum_{i=1}^{N} x_i^2\right) - N\bar{x}^2} = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad [12]$$

Volvamos ahora al ejemplo propuesto en la Tabla 1.a. En primer lugar calculamos en la Tabla 2 los valores de los elementos que forman las ecuaciones normales que integran el sistema expresado en [9], así como los valores medios de la variable explicada GAV y de la variable independiente RDH.

	A	B	C	D	E
1	obs.	GAV	RDH	RDH · GAV	$(RDH)^2$
2	1	9808	20022	196381581	400896020
3	2	8084	14854	120074067	220643927
4	3	5862	12009	70395212	144222344
5	4	10153	21139	214635686	446861119
6	5	9586	20012	192036671	400476704
7	6	10640	19326	205620540	373486765
8	7	6673	11503	64101904	132317032
9	8	8953	16293	145861483	265453449
10	9	9803	17563	172164645	308454990
11	10	9449	15418	145677389	237700180
12	11	5359	9996	51263084	91906207
13	12	9606	13820	117553461	190995234
14	13	4620	8263	38825491	68269084
15	14	11104	18499	205405668	342199944
16	15	9058	14112	130644406	199151879
17	16	7831	15050	117851258	226500531
18	17	7194	13667	98319963	186795084
19	18	7735	12796	98984639	163747463
20	19	9196	14140	130034324	199926868
21	20	7695	11586	87991529	134226164
22					
23	**Total:**	166617	299637	2606825280	4733837486
24					
25	**Media:**	8330,85	14981,84		
26					

TABLA 2

Así, para el ejemplo propuesto, la resolución del sistema

$$\begin{aligned} 166617 &= 20\hat{\alpha} + 299637\hat{\beta} \\ 2606825280 &= 299637\hat{\alpha} + 4733837486\hat{\beta} \end{aligned}$$

nos ofrece los coeficientes que definen la recta de regresión buscada: $\hat{\alpha}$ =1560,3335, $\hat{\beta}$ = 0,451914937. Por tanto, la estimación por MCO del modelo propuesto en [4] para explicar el gasto de los hogares en alimentación y vestido a partir de su renta disponible es:

$$\widehat{GAV} = 1560{,}3335 + 0{,}451914937 \cdot RDH \qquad [13]$$

Véamos ahora qué capacidad explicativa tiene el modelo estimado mediante la incorporación de la recta definida en [13] al Gráfico 1.

	A	B	C	D	E
1	obs.	RDH	GAV	GÂV	$\hat{\varepsilon}$
2	1	20022	9908	10608,78	-700,83
3	2	14864	8084	8273,12	-189,55
4	3	12009	5862	6987,50	-1125,75
5	4	21139	10153	11113,40	-959,91
6	5	20012	9596	10604,02	-1007,80
7	6	19326	10640	10293,98	345,68
8	7	11503	5573	6758,67	-1186,01
9	8	16293	8953	8923,27	29,28
10	9	17563	9803	9497,26	305,49
11	10	15418	9449	8527,74	921,07
12	11	9566	5359	5883,30	-524,32
13	12	13820	8606	7805,84	700,13
14	13	8263	4820	5294,29	-474,26
15	14	18499	11104	9920,14	1183,69
16	15	14112	9258	7937,81	1319,79
17	16	15050	7831	8361,62	-530,94
18	17	13667	7194	7736,79	-542,98
19	18	12796	7735	7343,21	392,15
20	19	14140	9196	7950,24	1246,21
21	20	11586	7595	6796,04	798,86
22					
23	**Total:**	299637	166617		0,00
24					
25	**Media:**	14981,84	8330,85	8330,85	
26					

FIGURA 3

Para ello, primeramente es necesario conocer los valores de la variable dependiente que proporciona la estimación muestral del modelo estocástico propuesto en la ecuación [5] para cada observación de la variable explicativa. De este modo se incorporan a la hoja de cálculo mostrada en la Tabla 3 los valores de $\widehat{GAV}_i$. Los residuos obtenidos como diferencia entre los valores observados y estimados de la variable dependiente $(GAV_i - \widehat{GAV}_i)$ constituyen, como ya se dijo, la estimación de la perturbación aleatoria para cada observación. Estos valores e_i de son calculados en la columna E de la Tabla.

Empleando el Asistente para Gráficos de Excel, ampliamos ahora el rango de selección de las observaciones a la variable $\widehat{GAV}_i$ (columna D). Para trazar la recta de regresión, hacemos doble *click* sobre cualquiera de los puntos marcados para la serie $\widehat{GAV}_i$, y una vez abierta la ventana de «Formato de serie de datos», eligimos la pestaña «Tramas», en la que procedemos a activar la opción de «Línea», personalizando con las correspondientes alternativas de Estilo, Color y Grosor para que destaque la línea en el gráfico. Si queremos que también se muestren los puntos correspondientes a los valores estimados por la recta, debemos también activar la opción de «Marcador».

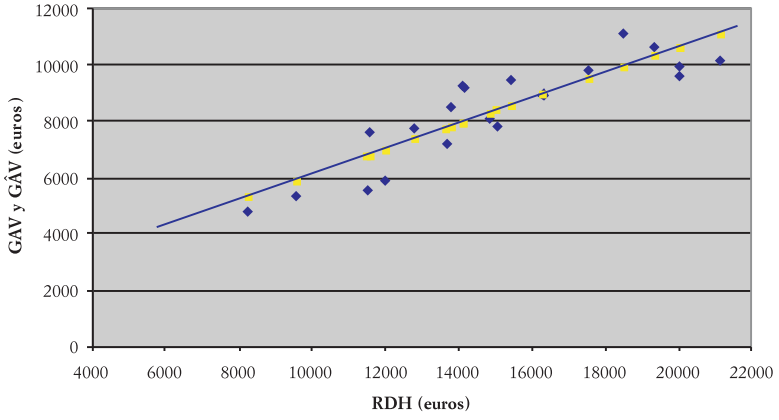


GRÁFICO 2. *Estimación MCO de la relación lineal entre Gastos en Alimentacion y Vestido - Renta Disponible del Hogar (2005)*

En la Tabla 3, las celdas C25 y D25 muestran que las medias de la variable *GAV* y de su estimación $\widehat{GAV}$ son iguales. Como señalamos al presentar la expresión [10], esto permite comprobar que la recta de regresión pasa por el punto correspondiente al par formado por los valores medios de ambas variables (14981,84; 8330,85).

El modelo de regresión lineal multiple

A diferencia del modelo de regresión lineal simple, el *modelo de regresión lineal múltiple*, también conocido como modelo de regresión lineal de *k* variables, extiende la relación lineal entre la variable dependiente o explicada *y* a dos o más (*k*) variables independientes o explicativas, x_1, x_2,..., x_k . Se trata, por tanto, de una generalización del modelo de regresión lineal, en el que el modelo simple sería un caso particular con una única variable explicativa. Por tanto, la ecuación que recoge este modelo lineal con *k* variables independientes:

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_k x_{ki}+\varepsilon_i \qquad [14]$$

La formulación presentada nos muestra la ecuación relacional en términos estocásticos, siendo válidos los argumentos expuestos para el modelo lineal simple. La notación empleada en la expresión [14] es consiste con el desarrollo matricial del modelo expuesto en la expresión [2]. La generalización de la relación lineal para *k* variables explicativas únicamente afecta a la dimensión de la matriz **X**, que ahora pasa a ser de orden $N\times(k+1)$, puesto que las observaciones de la variable dependiente han de seguir recogiéndose en un vector columna de dimensión $N\times 1$:

$$\mathbf{y}=\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_{N-1} \\ y_N \end{pmatrix};\ \mathbf{X}=\begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1N-1} & x_{2N-1} & \cdots & x_{kN-1} \\ 1 & x_{1N} & x_{2N-1} & \cdots & x_{kN-1} \end{pmatrix} \qquad [15]$$

Como en la expresión [3], en la matriz **X** se incorpora el término constante b_0 como una variable explicativa más, lo que supone que todos los elementos de su primera columna sean unos. Hay que señalar que para que el problema de regresión tenga una solución satisfactoria es necesario que la matriz **X** tenga *rango completo*, lo que equivale a decir que su rango es

igual al número de columnas que contiene $(k+1)$. Esta exigencia garantiza que no existe dependencia lineal entre las columnas de **X**. Si no fuese así, al menos una de las columnas sería una combinación lineal del resto de columnas.

De igual forma que se hizo para el modelo lineal simple, ilustraremos el desarrollo del modelo de regresión lineal múltiple con un ejemplo (ejemplo 2). En la Tabla 4 incluimos una serie de datos correspondientes a 20 observaciones de cuatro variables para trabajadores españoles en el año 2006: el salario anual (SAL), la edad del trabajador (EDAD), el sexo (GEN) y el nivel de estudios (EST). Tanto GEN como EST son variables cualitativas que toman valores por categorías. En el primer caso, 1 si el trabajador es hombre y 2, si es mujer, mientras que en la segunda variable, se contempla el valor 1 para trabajadores sin estudios, el 2 para estudios primarios, el 3 para estudios de enseñanza secundaria y el valor 4 para titulados superiores.

La estimación de «perfiles salariales» tanto transversales como longitudinales ocupa un lugar importante en los estudios de economía laboral. Dentro de este tipo de análisis, destacan recientemente los dedicados al estudio de la inversión en capital humano y, más en concreto, aquellos que analizan la relación existente entre los salarios y el nivel de formación de los trabajadores.

	A	B	C	D	E
1	**Obs.**	**SAL**	**EDAD**	**GEN**	**EST**
2	1	33889	49	1	4
3	2	21604	44	2	2
4	3	30186	36	1	3
5	4	34796	62	1	3
6	5	26917	40	2	2
7	6	29371	30	1	4
8	7	33733	54	1	3
9	8	24668	26	2	3
10	9	30089	56	1	1
11	10	29780	52	2	2
12	11	28024	29	2	4
13	12	31498	46	1	3
14	13	27679	38	2	2
15	14	33678	45	1	4
16	15	20775	19	1	1
17	16	20653	25	2	1
18	17	26989	39	1	3
19	18	37939	64	1	4
20	19	25116	32	2	2
21	20	30575	46	1	2
22					
23					
24	Sumx:	580657	830	26	53
25					
26					
27	Media:	29042,85	41,50		
28					

FIGURA 4

Con los datos recogidos en la Tabla 4 proponemos el siguiente modelo lineal consistente con los postulados teóricos señalados,

$$SAL=\beta_0+\beta_1\cdot EDAD+\beta_2\cdot SEXO+\beta_3\cdot EST \qquad [16]$$

[16]

donde el parámetro constante β_0 representaría la ordenada en el origen de la ecuación de regresión, mientras que los parámetros β_1, β_2 y β_3, y son las pendientes respectivas que recogen respectivamente cómo influye en el salario percibido por el trabajador su edad, si es hombre o mujer y su nivel de estudios. Los coeficientes de estos cuatro parámetros son las incógnitas a resolver en el proceso de estimación del modelo. Al igual que sucedía en el modelo de consumo formulado en la expresión [5], resulta necesario incorporar la aleatoriedad que sin duda existe en el mundo real a la hora de la determinación de los salarios. De nuevo, redefinimos la ecuación [16] en términos estocásticos, añadiendo el *shock* aleatorio de forma aditiva:

$$SAL=\beta_0+\beta_1\cdot EDAD+\beta_2\cdot SEXO+\beta_3\cdot EST+\varepsilon \qquad [17]$$

Regresión por Mínimos Cuadrados Ordinarios (MCO) del modelo lineal múltiple

Para realizar la regresión por MCO del modelo lineal de *k* variables se procede de igual forma que en el caso del modelo lineal simple. Para estimar los valores de los parámetros $\beta_0,\beta_1,\beta_2,\ldots,\beta_k$ que intervienen en la especificación de la relación lineal genérica expuesta en [14] debemos resolver el sistema de ecuaciones normales formado por las *k* condiciones de primer orden del problema de optimización que aseguran que la suma de los cuadrados de los residuos es la mínima factible. En forma matricial, este sistema de ecuaciones será del tipo [2]:

[2] Esta especificación es válida cuando la primera variable explicativa es una constante, y por tanto, la primera columna de la matriz X está formada por valores 1, como vimos. En caso contrario, se suprimiría la primera fila y la primera columna de la matriz **X**, así como la primera fila de los vectores $\mathbf{b}$ e **y.**

$$\begin{pmatrix} N & \sum_{i=1}^{N} x_{1i} & \sum_{i=1}^{N} x_{2i} & \cdots & \sum_{i=1}^{N} x_{ki} \\ \sum_{i=1}^{N} x_{1i} & \sum_{i=1}^{N} x_{1i}^2 & \sum_{i=1}^{N} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{N} x_{1i}x_{ki} \\ \sum_{i=1}^{N} x_{2i} & \sum_{i=1}^{N} x_{2i}x_{1i} & \sum_{i=1}^{N} x_{2i}^2 & \cdots & \sum_{i=1}^{N} x_{2i}x_{ki} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{N} x_{ki} & \sum_{i=1}^{N} x_{ki}x_{1i} & \sum_{i=1}^{N} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{N} x_{ki}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{1i}y_i \\ \sum_{i=1}^{N} x_{2i}y_i \\ \cdots \\ \sum_{i=1}^{N} x_{ki}y_i \end{pmatrix} \quad [18]$$

Los valores de los elementos que definen estas ecuaciones normales resultan del desarrollo de la siguiente identidad en la que intervienen las dos matrices de rango completo, **X** (formada por la totalidad de datos observados para las variables explicativas, incluido el término constante) e **y** (integrada por los datos de la variable explicada para esas mismas observaciones):

$$(\mathbf{X}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad [19]$$

La solución se obtiene despejando en el sistema el vector $\hat{\boldsymbol{\beta}}$, lo que nos permite obtener los coeficientes correspondientes al término constante y a las *k* variables explicativas del modelo. Operando matricialmente a partir de [19], la solución será:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad [20]$$

donde **X'** es la matriz traspuesta de **X, y** $(\mathbf{X}'\mathbf{X})^{-1}$ es el resultado de invertir el producto de **X'X**.

Los valores de las incógnitas calculados y recogidos en `b` son la estimación MCO del vector de parámetros que intervienen en el modelo formulado en [14]. Es decir, `b` es el estimador MCO de `b`.

De igual forma que sucedía en el modelo con una única variable explicativa, el vector de estimadores MCO nos permite calcular para las N observaciones las correspondientes estimaciones de la variable dependiente, $\hat{y}_i$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki} \quad [21]$$

En la medida que $\hat{y}_i$ es una estimación del verdadero valor que se observaría en realidad para y_i, del proceso de regresión se obtiene para cada una de las observaciones un residuo, $e_i = y_i - \hat{y}_i$, que como ya expusimos en el caso del modelo lineal simple constituye el estimador de la correspondiente perturbación aleatoria. Por tanto, para las N observaciones tendremos un vector de residuos de dimensión $N \times 1$:

$$e = y - \hat{y} = y - X\hat{\beta} \qquad [22]$$

Comprobemos ahora que el vector de estimadores MCO $\hat{\beta}$ conduce al mejor ajuste lineal posible para los datos de la muestra bajo el criterio de minimización de la suma de los residuos al cuadrado. Si definimos matricialmente la suma de los N residuos al cuadrado de acuerdo con [22], tal que,

$$\rho(\beta) = \sum_{i=1}^{N} e_i^2 = e'e = (y - X\beta)'(y - X\beta) = y'y - 2\beta'X'y + \beta'X'X\beta \qquad [23]$$

el método de regresión MCO deben asegurar que el valor $\rho(\hat{\beta})$ obtenido para el vector solución del sistema de ecuaciones [18] es mínimo frente a cualquier otro conjunto de coeficientes estimadores del vector de parámetros b. Por tanto, resolvemos el problema de optimización,

$$\min_{\beta} \rho(\beta) = \min_{\beta} (y'y - 2\beta'X'y + \beta'X'X\beta) \qquad [24]$$

obteniendo la siguiente condición necesaria de mínimo:

$$\text{CPO}(\hat{\beta}): \quad \frac{\partial \rho(\beta)}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0 \qquad [25]$$

Reagrupando términos, formulamos el sistema de ecuaciones normales $X'X\hat{\beta} = X'y$, coincidente con el expuesto en forma desarrollada en [18]. Por consiguiente, tenemos que el vector de estimadores MCO $\hat{\beta}$, que de acuerdo con [20] es la solución de este sistema, satisface la condición de hacer mínima la suma de los N residuos elevados al cuadrado. Para asegurar este resultado, comprobamos que la matriz hessiana formada por las segundas derivadas es «definida positiva». Sea,

$$\frac{\partial^2 \rho(\beta)}{\partial \beta \partial \beta'} = 2X'X \qquad [26]$$

En algebra matricial es un resultado demostrado que para un vector cualquiera distinto de cero. Únicamente tendríamos que **c'X'Xc**=0 si para todas las observaciones se verificase que . Sin embargo, si esto fuese así, es inmediato que tendría que ser una combinación lineal de columnas de la matriz que se igualarían a cero, lo que se opondría a la exigencia de *rango completo* establecida anteriormente para **X**. Se demuestra así que el vector MCO $\hat{\beta}$ es la solución que hace mínima la suma de los cuadrados de los residuos.

Una vez expuesto el método de obtención de la regresión MCO para el modelo lineal general, vamos a aplicarlo a la estimación de la ecuación de salarios propuesta en [17]. Comencemos formulando las ecuaciones normales que forman el sistema a resolver, de acuerdo con su desarrollo matricial expuesto en [18]. Para ello previamente es necesario obtener a partir de los datos recogidos en la Tabla 4 todos los elementos intervienen en [18]. Sus valores calculados (las sumas de las correspondientes columnas) se muestran en la Tabla 5.

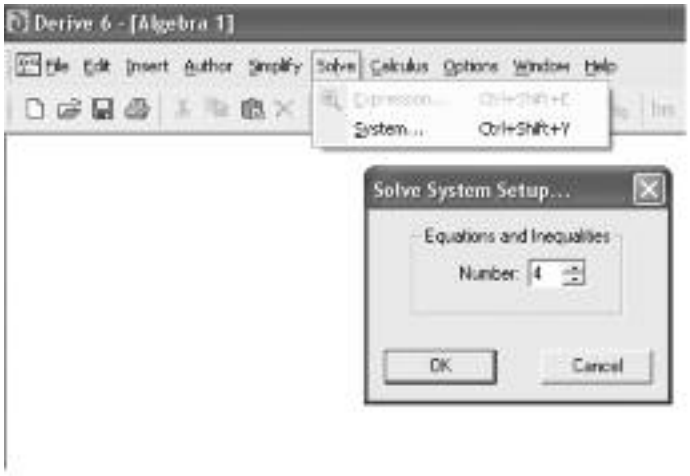
	A	B	C	D	E	F	G	H	I	J
1	Obs.	EDAD	GEN	EST	EDAD x GEN	EDAD x EST	GEN x EST	EDAD x SAL	GEN x SAL	EST x SAL
2	1	2401	1	16	49	196	4	1680657	83887	135556
3	2	1369	4	4	86	86	4	2545678	43308	43308
4	3	1296	1	9	36	108	3	1889889	36186	30667
5	4	3844	1	9	62	186	3	2757373	34706	104389
6	5	1600	4	4	80	80	4	1078880	53931	53931
7	6	900	1	16	30	120	4	881121	29371	117483
8	7	2916	1	9	54	162	3	1871670	33703	101199
9	8	676	4	9	52	78	6	641368	49336	74004
10	9	3025	1	1	55	55	1	1992800	36389	36389
11	10	2774	4	4	104	104	4	1589950	59998	59998
12	11	841	4	16	29	116	4	613696	99540	112086
13	12	2116	1	9	46	138	3	1489975	31490	94495
14	13	1444	4	4	76	76	4	1099402	55790	55790
15	14	2025	1	16	45	180	4	1510930	33576	134305
16	15	361	1	1	19	19	1	394722	20775	20775
17	16	625	4	1	50	25	2	521325	41706	20853
18	17	1444	1	9	38	114	3	1101584	28989	86967
19	18	4096	1	16	64	256	4	2428102	37939	151756
20	19	1024	4	4	64	64	4	803712	50232	50232
21	20	2116	1	4	46	92	2	1424843	30975	61950
22										
23										
24	Suma	39330	44	161	1116	2257	71	24996636	785698	1599265
25										

FIGURA 5

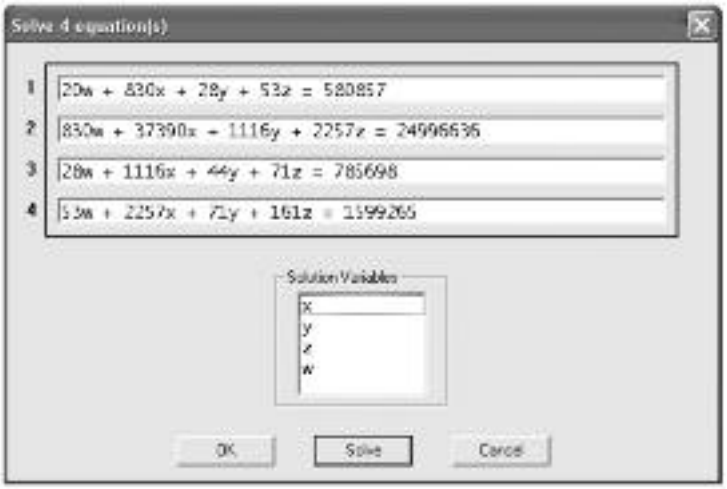
Así, el sistema de ecuaciones a resolver para obtener los coeficientes de $\hat{\beta}$ será:

$$\begin{pmatrix} 20 & 830 & 28 & 53 \\ 830 & 37390 & 1116 & 2257 \\ 28 & 1116 & 44 & 71 \\ 53 & 2257 & 71 & 161 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} 580857 \\ 24996636 \\ 785698 \\ 1599265 \end{pmatrix} \qquad [27]$$

Resolvemos este sistema de ecuaciones utilizando el programa Derive^{TM}6. Seleccionando la función Solve, elegimos la opción System, y en la ventana que aparece y que mostramos definimos un sistema formado por cuatro ecuaciones:



A continuación, mediante el asistente para la resolución de sistemas de ecuaciones, tecleamos las cuatro ecuaciones normales de [26] en su forma ordinaria. Y elegimos las incógnitas a despejar $w=\hat{\beta}_0, x=\hat{\beta}_1, y=\hat{\beta}_2, z=\hat{\beta}_y$



Haciendo click sobre el botón Solve, obtenemos los resultados del sistema. Para expresar los valores en números reales aproximados, hay que marcar en la barra de herramientas de Derive 6 la opción Simplify, y luego «Approximate»):

x = 229,9379800; y = –2241,284680; z =1927,027015; w = 17531,60078

Calculados los coeficientes de cada variable explicativa, la ecuación de regresión correspondiente a la estimación del modelo lineal propuesto en la expresión [17] será:

$$SAL = 17531{,}60 + 229{,}938 \cdot EDAD - 2241{,}285 \cdot GEN + 1927{,}027 \cdot EST$$

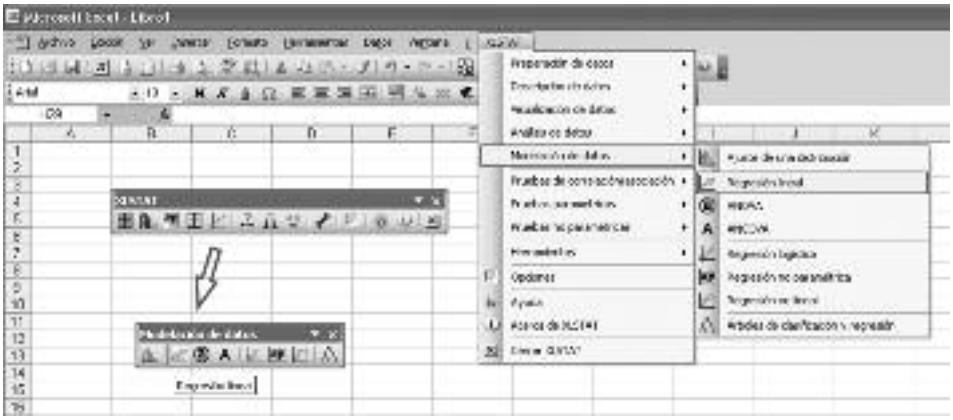
15.3. ESTIMACIÓN MCO CON XLSTAT 2007 DEL MODELO DE REGRESIÓN LINEAL

Si el modelo de regresión lineal incorpora más de tres variables explicativas (o lo que es lo mismo, la matriz **X** el rango es superior a 3), la resolución del sistema de ecuaciones normales que permite obtener los coeficientes de $\hat{\beta}$ se convierte en una tarea bastante complicada. Como vimos en el ejemplo 2 una alternativa es utilizar *software* matemático para operar matricialmente y resolver [20]. Sin embargo, la generalización de los métodos econométricos ha promovido la producción de *software* específico para el análisis econométrico de fácil disposición en el mercado. Estos programas no sólo permiten la obtención directa a partir de grandes bases de datos de los coeficientes de regresión, sino que además aportan la mayoría de los tests y contrastes estadísticos que permiten valorar la bondad de las estimaciones realizadas.

En este capítulo presentamos la aplicación del programa XLSTAT 2007 de Addinsoft (http://www.addinsoft.com) dirigida a la realización de análisis de regresión lineal. XLSTAT 2007 es un paquete de análisis estadístico de datos cuya principal potencialidad es su fácil manejo desde las hojas de cálculo del programa Excel de Microsoft (en cualquiera de sus versiones, desde el Excel 97 hasta el Excel 2007). De hecho, el acceso al programa se realiza a través de Excel. Después de la instalación de XLSTAT 2007, en la barra de herramientas de Excel aparece el icono que da acceso al programa. Al hacer click en este icono, aparece en pantalla la ventana de advertencia que nos interroga acerca de la seguridad de las «macros» que contiene el programa de Addinsoft (la programación de XLSTAT 2007 está realizada lenguaje C++, y en muchos casos operan como «macros» de Excel), y elegimos la opción «Habilitar macros». tal y Como vemos en la siguiente figura, inmediatamente en la hoja de cálculo abierta

aparecen una nueva barra de herramientas, «XLSTAT» y el icono del programa XLSTAT 2007 se activa, (para salir del programa basta con volver a hacer click sobre él). Para utilizar todas las aplicaciones contenidas en el programa XLSTAT 2007 se puede acceder a través de los iconos contenidos en la barra «XLSTAT». Cada vez que se elige uno de estos iconos, automáticamente nos aparece sobre la hoja de cálculo una nueva barra con todos los iconos correspondientes a las herramientas y funciones que el programa incluye para esa aplicación concreta. Así, como podemos ver en la figura que sigue, al hacer click en el icono se muestra la barra de herramientas de «Modelación de datos», con sus distintos iconos.

No obstante, observamos que en la barra del menú principal de Excel aparece, tras la entrada al menú de ayuda ?, una entrada a un nuevo menú, XLSTAT, que mediante menús desplegables permite seleccionar también la totalidad de las aplicaciones de este programa, cada una de ellas con sus correspondientes funciones y herramientas.

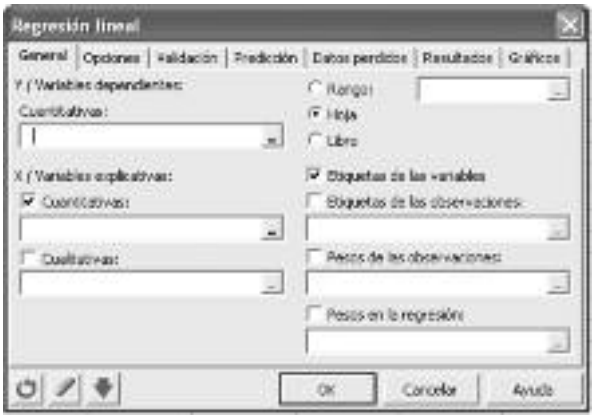


Entre las herramientas econométricas para el análisis de datos que incorpora el programa XLSTAT 2007 se encuentra un módulo específico de regresión lineal. Como vemos en la figura, para acceder al mismo, bien se puede entrar a través del menú de «Modelación de datos», o bien marcando en el icono que aparece en la barra asociada a este menú.

Para ilustrar el procedimiento a seguir en la realización de regresiones lineales con XLSTAT 2007 resolvemos nuevamente el ejercicio de estimación de un modelo de salarios a partir de los datos que figuran en la Tabla 4. Como hemos dicho, XLSTAT 2007 trabaja en sus distintas aplicaciones como si se tratase de herramientas propias de Excel. Por tanto, el primer paso es disponer en una hoja de cálculo los datos que se utilizarán en el análisis de regresión. Para ello recuperamos la Tabla 4, en la que fi-

guraban por columnas los datos correspondientes a las variables que intervenían en el modelo propuesto en la expresión [16], «salario» (SAL), «edad» (EDAD), «sexo» (GEN) y «nivel de estudios» (EST).

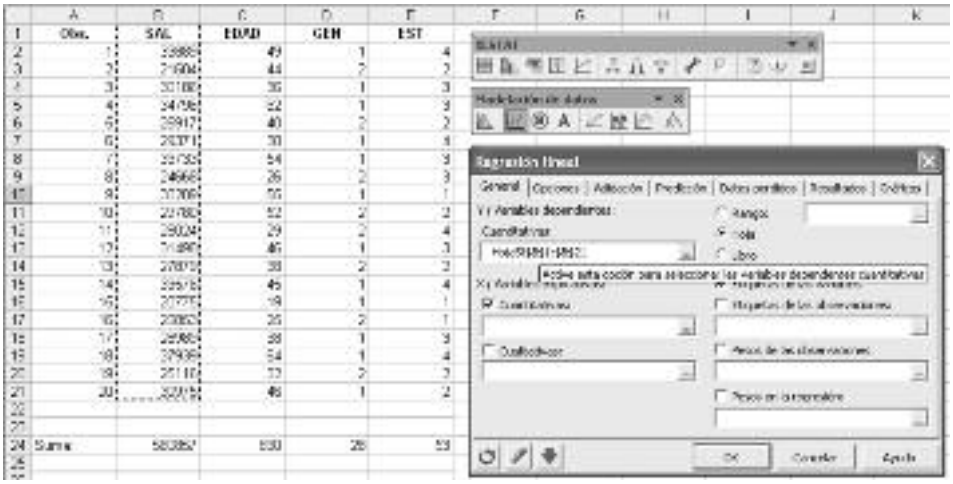
Una vez tenemos los datos presentados para el análisis, el segundo paso es abrir la herramienta de «Regresión lineal» de XLSTAT 2007, lo que hacemos marcando en el icono . De inmediato aparece en pantalla el asistente para aplicar esta herramienta. Para comenzar a introducir los datos, elegimos la ventana «General», que mostramos a continuación.

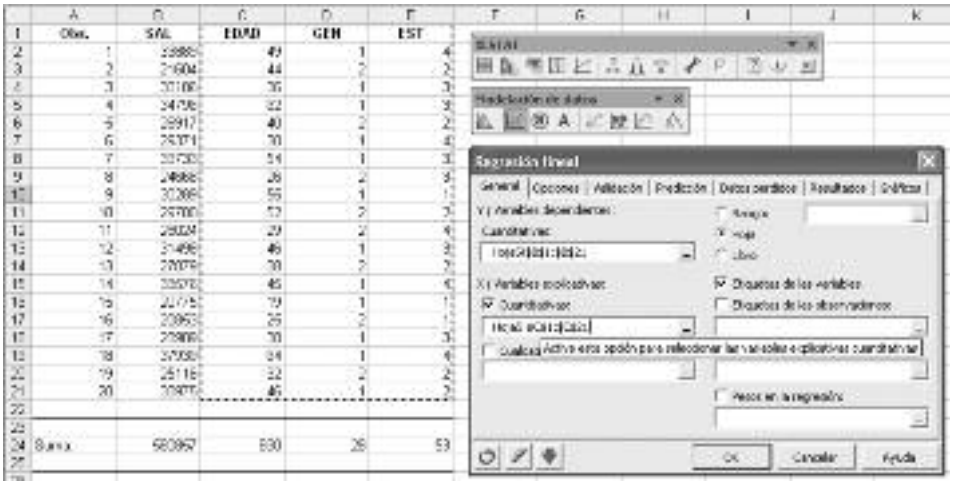


En primer lugar, señalamos a través de la opción «Etiquetas de las variables» si las series de datos recogidas en la Tabla 4 tienen o no encabezados con el nombre de la variable. Es recomendable usar esta opción, pues al obtener los resultados de la regresión, las referencias a cada variable aparecen con el nombre que figura en el encabezado de la columna correspondiente. De lo contrario, el programa define a cada variable como Obs1, Obs2, etc. No hay que olvidar que esto requiere aumentar la selección de celdas a las casillas que contienen los nombres de las variables. También es necesario elegir la categoría de las variables explicativas (Cuantitativas o Cualitativas), si bien lo habitual será usar la opción por defecto (Cuantitativas), al estar integrada la matriz **X** por variables numéricas [3]. La opción «Etiquetas de las observaciones»: requiere la selección de los datos correspondientes que, en caso de existir, contendrían la información de cada observación. No es frecuente su utilización en las bases de datos y, por tanto, lo normal es dejar sin marcar esta opción.

[3] La opción «Cualitativas» es adecuada para la realización de un análisis ANCOVA, con datos seleccionados no cuantitativos.

Para incorporar los datos al análisis, en primer lugar situamos el cursor en la entrada de datos correspondiente a la variable dependiente, señalada en el asistente como «Y / Variables dependientes». A continuación seleccionamos en la hoja de cálculo el bloque correspondiente a la columna en la que se contienen los datos de la variable dependiente, incluido el encabezado. En nuestro ejemplo, seleccionamos el rango $B$1:$B$21, que queda incorporado en el asistente, tal y como vemos en la siguiente figura. A continuación, introducimos con el mismo procedimiento las variables explicativas, aunque en este caso se selecciona en conjunto el rango definido por las columnas de todas esas variables y por las observaciones de cada una de ellas. En nuestro caso, seleccionamos el rango $C$1:$E$21. Si en la muestra de datos, las observaciones tuviesen pesos poblacionales diferenciados, deberíamos seleccionar el vector correspondiente a través de la opción «Pesos de las observaciones». Por defecto, el peso es la unidad.





El asistente ofrece la posibilidad de invertir la presentación de los datos, intercambiando filas y columnas. Para ello, puede utilizarse alternativamente el botón (variables en columnas, observaciones en filas) o el botón (variables en filas, observaciones en columnas).

Una vez introducidos en el asistente los datos de todas las variables que intervienen en la ecuación a estimar, hacemos click en el botón OK para ejecutar la aplicación que realiza la regresión lineal. Si tenemos activada la opción de «mostrar selección, aparecerá esta ventana con la selección de variables realizada.

Comprobado que la selección es correcta, marcamos en el botón Continuar. De inmediato aparece una nueva hoja de cálculo (con el nombre «Regresión lineal») que incorpora tanto los resultados de la estimación como un completo análisis estadístico de la regresión realizada. Por el momento, nos detendremos en el cálculo de los coeficientes del vector β. Como vemos en la siguiente figura, los valores de los parámetros del modelo coinciden con los obtenidos de la resolución del sistema de ecuaciones normales [27]. Las diferencias en los decimales son consecuencia de los diferentes algoritmos de resolución del sistema de ecuaciones empleados.

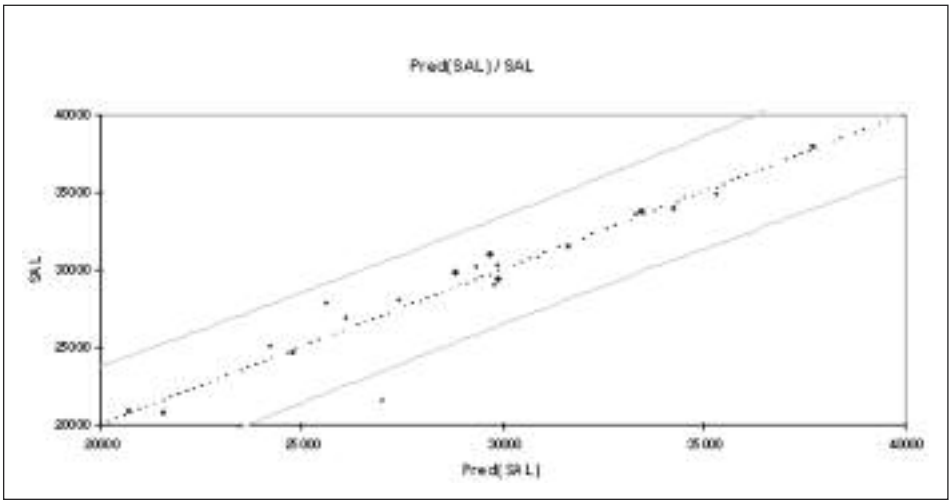
	A	B	C	D	E	F	G	H
53								
54		Parámetros del modelo						
55								
56		Fuente	Valor	Desviación típica	t	Pr > |t|	Límite inferior (95%)	Límite superior (95%)
57		Intersección	17531,631	2369,807	7,399	< 0,0001	12907,610	22999,652
58		EDAD	220,937	32,498	7,075	< 0,0001	161,045	288,829
59		GEN	-2241,282	806,256	-2,711	0,015	-3863,359	-488,647
60		EST	1927,036	370,949	6,886	0,000	1123,701	2730,372
61								
62								
63		Ecuación del modelo						
64								
65		SAL = 17631,631+220,937*EDAD-2241,282*GEN+1927,036*EST						
66								

Además de los coeficientes, en la hoja de cálculo de «Regresión lineal» se muestra una tabla de estadísticas básicas de las variables, así como su matriz de correlaciones:

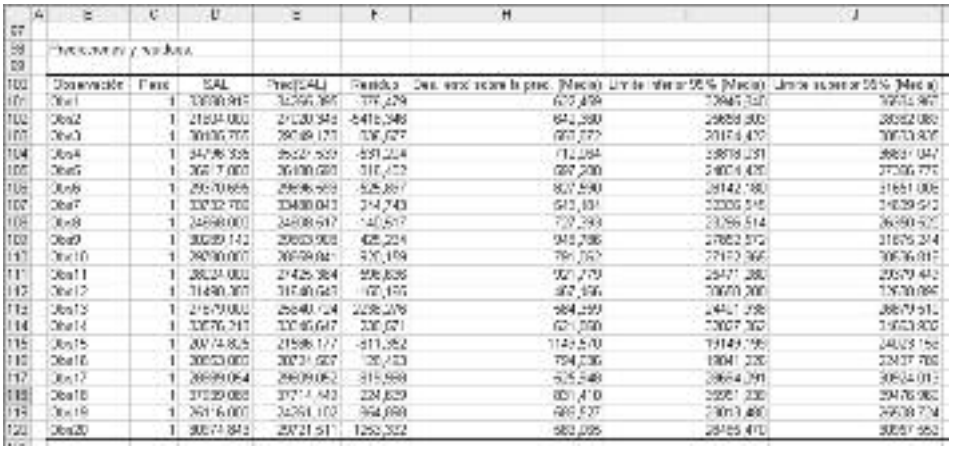
	A	B	C	D	E	F	G	H	I
7									
8		Estadísticos simples							
9									
10		Variable	Observaciones	Obs. con datos perdidos	Obs. sin datos perdidos	Mínimo	Máximo	Media	Desviación típica
11		SAL	20	0	20	20174,825	37239,288	28342,662	4722,812
12		EDAD	20	0	20	19,000	64,000	41,500	12,460
13		SEN	20	0	20	1,000	2,000	1,400	0,503
14		EST	20	0	20	1,000	4,000	2,950	1,040
15									
16									
17		Matriz de correlación							
18									
19		Variables	EDAD	SEN	EST	SAL			
20		EDAD	**1,000**	-0,287	0,254	0,798			
21		SEN	-0,287	**1,000**	-0,502	-0,610			
22		EST	0,254	-0,502	**1,000**	0,643			
23		SAL	0,798	-0,610	0,643	**1,000**			
24									

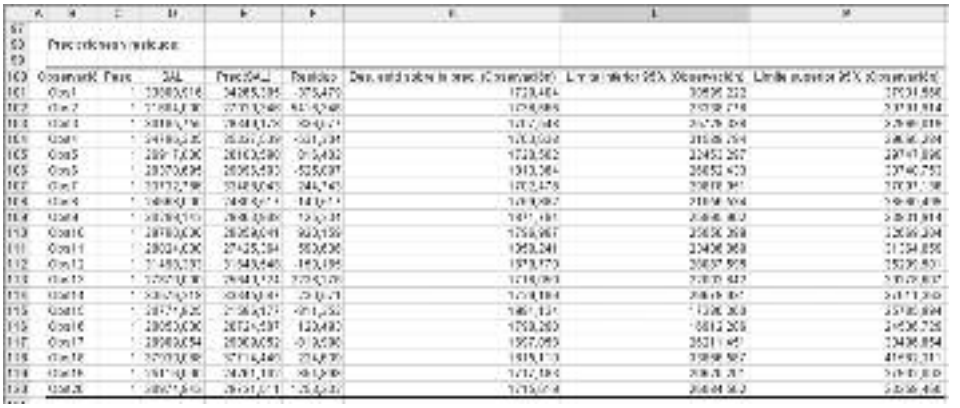
En la matriz de correlaciones podemos observar la elevada correlación positiva existente entre el salario y la edad de los trabajadores (0,798), y algo menor entre el salario y el nivel de estudios (0,643). El valor negativo de la correlación entre salario y sexo, dado el valor dicotómico de esta variable explicativa (1 para hombres, 2 para mujeres), muestra la existencia de un diferencial favorable al salario de los hombres.

La aplicación ofrece el cálculo de los residuos de la regresión, como diferencia entre el valor de cada observación de la variable dependiente, SAL, y el valor que predice la ecuación estimada, Pred(SAL). También se muestra gráficamente la relación entre ambos valores de cada observación.



Además, la tabla de residuos recoge para cada observación el valor de la desviación estándar tanto sobre la media como sobre la predicción, con los correspondientes límites superior e inferior que definen el intervalo para un nivel de confianza del 95% [4].





15.4. VALORACIÓN DE LOS RESULTADOS DE LA ESTIMACIÓN MCO

Hasta ahora, el proceso de regresión del modelo lineal ha sido presentado como la resolución de un problema estrictamente algebraico, en el que los parámetros del modelo estimado se identificaban con las incógnitas del problema planteado y el método de estimación MCO constituía el método para su resolución. Sin embargo, una vez obtenidos los resultados de la re-

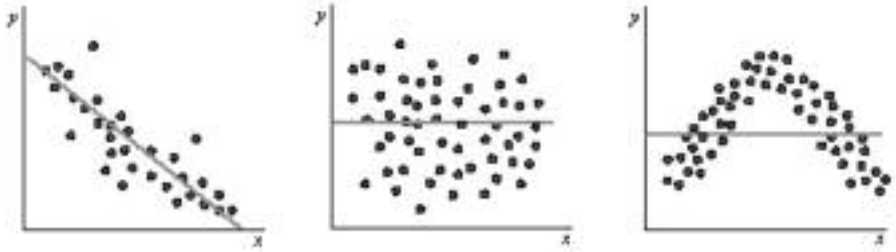
[4] Para modificar el intervalo de confianza, fijado por defecto en el 95%, hay que cambiar el valor en la pestaña «Opciones» del asistente de «Regresión lineal».

gresión es necesario realizar una valoración estadística de los mismos. Se trata de aceptar o rechazar la capacidad del modelo estimado para explicar la proposición teórica que subyace al mismo, así como su capacidad predictiva. Este proceso de evaluación se ha de abordar desde una doble aproximación. Por un lado, hay que validar esas potencialidades del modelo estimado considerándolo en su conjunto. Por otro, resulta necesario contrastar el poder explicativo de cada una de las variables que intervienen en el mismo. La econometría nos ofrece un conjunto de instrumentos estadísticos para poder realizar una valoración normalizada de cualquier estimación. A continuación presentamos los principales, disponibles en la práctica totalidad de los paquetes informáticos para el análisis econométrico.

Bondad del ajuste

Para analizar la precisión de un estimador suelen utilizarse El método de estimación MCO incorpora un criterio de ajuste razonable para llevar a cabo la regresión de un modelo lineal a partir de un conjunto de observaciones: hacer mínima la suma de los cuadrados de los residuos resultantes. Sin embargo, el cumplimiento de este criterio sobre el que descansa el procedimiento de obtención de los coeficientes de regresión MCO no garantiza la *bondad del ajuste*. No debemos olvidar que la labor del investigador se inicia con la propuesta de un modelo en el que una combinación lineal de diversas variables explicativas o exógenas se supone capaz de determinar el valor de una variable dependiente. La recopilación de datos procedentes de las observaciones en N instantes o casos de los valores para esas k variables explicativas y para la variable explicada configuran el conjunto de datos muestrales a partir de los cuales obtenemos los estimadores MCO que caracterizan esa relación lineal. Obviamente, su obtención, haciendo mínima la suma de los residuos al cuadrado, no garantiza que la relación entre variables propuesta en el modelo teórico sea la adecuada. En otras palabras, *a priori* desconocemos el potencial explicativo del modelo estimado.

Para ilustrar esta consideración, podemos presentar varias estimaciones del modelo de regresión lineal simple correspondientes a diferentes muestras de datos. Únicamente la figura de la izquierda recoge un ajuste a primera vista satisfactorio, en la medida que la disposición de los pares (x,y) de las observaciones muestran una relación que podría ser definida aceptablemente por la recta de regresión obtenida. En cambio, tanto en la figura de la derecha como en la del centro, la disposición de las relaciones (x,y) no permite encontrar un patrón que permita aceptar su explicación a través de la recta de ajuste estimada.



La introducción en el modelo de regresión lineal de dos o más variables explicativas, además del coeficiente constante, impide valorar gráficamente la idoneidad del ajuste realizado. Incluso, en el modelo lineal simple, la simple evaluación gráfica de la bondad del ajuste carece del rigor estadístico que exige la contrastación empírica de cualquier modelo teórico. La elección entre modelos alternativos obliga a encontrar un criterio estadístico acerca de la calidad del ajuste en cada caso.

La medida estadística habitualmente empleada para evaluar la bondad del ajuste MCO en el modelo de regresión lineal es el *coeficiente de determinación*. El coeficiente de determinación, generalmente representado como R^2, se define como el cuadrado de la correlación entre los valores observados de la variable explicada y los valores que para esa variable predice la ecuación de regresión estimada.

En términos matriciales, el valor del coeficiente de determinación se calcula como,

$$R^2 = \frac{\hat{\beta}' \mathbf{X}' \mathbf{M}^0 \mathbf{X} \hat{\beta}}{\mathbf{y}' \mathbf{M}^0 \mathbf{y}} = 1 - \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{\mathbf{y}' \mathbf{M}^0 \mathbf{y}} \qquad [28]$$

donde $\mathbf{M}^0$ es una matriz idempotente de rango $N \times N$ que transforma las observaciones en desviaciones respecto de las medias muestrales. En esta expresión podemos encontrar dos definiciones alternativas de R^2. Así, de acuerdo con el término de la izquierda, el coeficiente de determinación es el cociente entre la suma del cuadrado de las desviaciones de la variable estimada respecto de su media (ΣCR) y la suma del cuadrado de las desviaciones de la variable dependiente observada respecto de su media (ΣCT). El término de la derecha nos permite definir alternativamente R^2 como la diferencia entre la unidad y el cociente entre la suma del cuadrado de los errores (ΣCE) y la suma del cuadrado de las desviaciones de la variable dependiente observada respecto de su media (ΣCT). Por tanto, se cumple,

$$R^2 = \frac{\Sigma CR}{\Sigma CT} = 1 - \frac{\Sigma CE}{\Sigma CT} \qquad [29]$$

También resulta de utilidad la siguiente expresión de R^2 que utiliza directamente las sumas de las desviaciones respecto de la media tanto de los valores observados de y como de los predichos por el modelo estimado, $\hat{y}$:

$$R^2 = \frac{\left[\sum_{i=1}^{N}(y_i - \bar{y})\cdot(\hat{y}_i - \bar{\hat{y}})\right]^2}{\left[\sum_{i=1}^{N}(y_i - \bar{y})^2\right]\cdot\left[\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2\right]} \qquad [30]$$

En la Tabla 6 se presentan todos los valores necesarios para calcular el coeficiente de determinación de acuerdo con la expresión [30]. Los N valores estimados de la variable dependiente, recogidos en la columna SÂL han sido calculados a partir de la ecuación de regresión definida con los coeficientes obtenidos usando XLSTAT 2007.

Con estos datos, el valor del coeficiente de determinación correspondiente al modelo de regresión estimado es:

$$R^2 = \frac{(382139948{,}93)^2}{423794074{,}14 \times 382139948{,}93} = 0{,}901711$$

Además, comprobamos que si obtenemos la suma del cuadrado de los errores (ΣCE) y la suma del cuadrado de las desviaciones de la variable dependiente observada respecto de su media (ΣCT), llegamos al mismo valor,

$$R^2 = 1 - \frac{41654125{,}21}{423794074{,}14} = 0{,}901711$$

lo mismo que si es calculado a través del cociente entre ΣCR y ΣCT:

$$R^2 = \frac{382139948{,}93}{423794074{,}14} = 0{,}901711$$

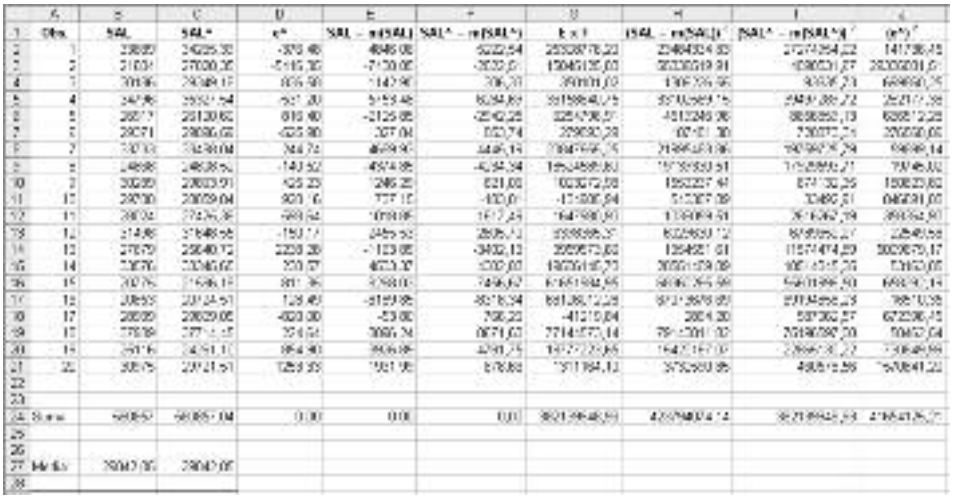


TABLA 6

Al reescribir [28] como $\mathbf{y'M^0y=\hat{\beta}'X'M^0X\hat{\beta}+\hat{\varepsilon}'\hat{\varepsilon}}$, obtenemos la siguiente relación, ΣCT=ΣCR+ΣCE, que podemos verificar que se cumple con los datos incluidos en la Tabla 6. Ante un ajuste de la regresión perfecto, tenemos que ΣCT=ΣCR, y por tanto, ΣCE=0, lo que hace que R^2=1. En cambio, si todos los coeficientes de $\hat{\boldsymbol{\beta}}$ salvo el término constante fuesen cero, la regresión se traduce en una recta horizontal con intersección en el valor $\hat{\beta}_0$ de la ordenada, lo que equivale a decir que **X** no tiene ningún poder explicativo respecto de **y**. En este caso, todos los valores estimados de la variable dependiente serán iguales a su media, $\hat{y}_i=\bar{y}$, y por tanto, ΣCR=0, lo que conduce a que R^2=0. Por consiguiente, tenemos que el coeficiente de determinación siempre estará comprendido entre 0 –cuando el modelo estimado muestra que no existe relación alguna entre **X** e **y**– y 1 –cuando la regresión muestra que **y** depende perfectamente, sin error alguno, de **X**.

La interpretación del coeficiente de determinación resulta bastante intuitiva. R^2 mide qué proporción (entre 0 y 100%) de la variación de la variable dependiente es explicada por las variaciones de las variables exógenas. En la medida que la influencia de estas variaciones de las variables explicadas en la variable dependiente se produce a través de los regresores, el coeficiente de determinación se convierte en un test para evaluar la capacidad predictiva del modelo, o lo que es lo mismo, en qué medida (porcentaje) el modelo estimado es capaz de explicar los cambios en la variable explicada.

A pesar de su sencillez e intuición, el coeficiente de determinación presenta algunos problemas. El principal tiene que ver con el número de

variables y el número de observaciones que intervienen en una regresión. A medida que introducimos una variable explicativa adicional en el modelo, tenemos que siempre el valor de R^2 va creciendo, hasta su límite en 1. Sin embargo, esto va agotando los grados de libertad del modelo (la diferencia entre el número de observaciones y el número de variables, esencial para una estimación robusta), lo que hace que los parámetros estimados vayan perdiendo en precisión.

Para paliar esta limitación, suele ofrecerse junto con el valor de R^2 un coeficiente alternativo denominado R^2 *ajustado*. Se trata de una modificación en la definición del R^2, basado en un ajuste en función de los grados de libertad ($N–k$) tal que,

$$\bar{R}^2 = 1 - \frac{N-1}{N-K}\left(1-R^2\right) \qquad [31]$$

En nuestro caso, el valor del coeficiente de determinación ajustado sería:

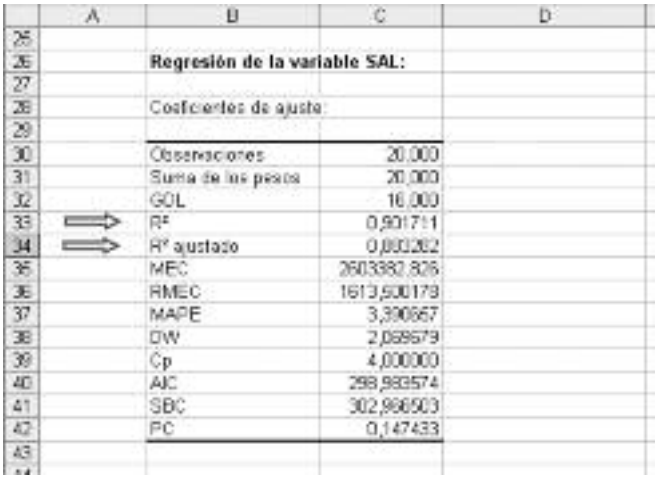
$$\bar{R}^2 = 1 - \frac{20-1}{20-4} \times \left(1-0{,}901711\right) = 0{,}883282$$

De acuerdo con [28], su expresión en términos matriciales es,

$$\bar{R}^2 = 1 - \frac{\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{N-k}}{\frac{\mathbf{y}'\mathbf{M}^0\mathbf{y}}{N-1}} \qquad [32]$$

No obstante, el coeficiente $\bar{R}^2$ también tiene algunas objeciones. La principal es que al añadir una variable explicativa más, al contrario que sucedía con R^2, su valor puede disminuir, incluso llegar a ser negativo (si $\mathbf{X}$ e $\mathbf{y}$ presentan una correlación muestral nula).

Todos los programas informáticos de análisis econométrico incluyen en sus resultados el cálculo de los coeficientes R^2 y $\bar{R}^2$. Volviendo a la resolución del Ejercicio 2 mediante el uso de XLSTAT 2007, esta información aparece en el primer bloque de resultados de la regresión:



	A	B	C	D
25				
26		**Regresión de la variable SAL:**		
27				
28		Coeficientes de ajuste:		
29				
30		Observaciones	20,000	
31		Suma de los pesos	20,000	
32		GDL	16,000	
33	⇒	R^2	0,901711	
34	⇒	R^2 ajustado	0,880282	
35		MEC	2603382,826	
36		RMEC	1613,500179	
37		MAPE	3,390667	
38		DW	2,089579	
39		Cp	4,000000	
40		AIC	298,983574	
41		SBC	302,966903	
42		PC	0,147433	
43				

Propiedades de los estimadores MCO

En el análisis econométrico suponemos habitualmente que las variables exógenas son no estocásticas. En la medida que en el mundo real los fenómenos con trascendencia económica no están sujetos a leyes deterministas, o lo que es lo mismo, no se producen en experimentos controlados de laboratorio, resulta muy difícil asumir esta suposición. Esta hecho resulta trascedental a la hora de evaluar las propiedades estadísticas de los parámetros estimados para los distintos regresores. Sin embargo, para poder aplicar el análisis estadístico convencional podemos movernos bajo la hipótesis de que las N observaciones de la variable explicativa x_k siempre se repiten para cualesquiera extracción muestral de dicha variable[5]. Así estaremos aceptando que el modelo de regresión es invariante ante el proceso muestral, pues los datos observados de x_k tendrán el carácter de constante cierta (no aleatoria).

Este supuesto es fundamental para garantizar que para cualquier observación, el valor esperado de la perturbación aleatoria ha de ser cero, $E[\varepsilon_i|\mathbf{X}]=0, \forall i=\{1,N\}$. Si no fuese así, es obvio que habría incluir este valor esperado (medio) de la perturbación en el modelo estimado, dejando fuera estrictamente su elemento estocástico (desconocido). Por consiguiente, este supuesto implica que la regresión de **y** sobre **X** a partir de los da-

[5] Esto no sucede con la variable dependiente, pues y es por definición aleatoria, al ser función del término de error, además de **X**. Por consiguiente, la extracción de distintas muestras de y supondría cambios en los valores observados. Este hecho no siempre se verifica, pues en series temporales de tipo agregado es lógico imaginar que los valores de y para cada año (p. e. la presión fiscal o el gasto en educación/PIB) son únicos. En cambio, sí se producirá ante la selección de distintas muestras de microdatos (p. e. el consumo anual de los hogares o el salario de los adultos de un hogar).

tos recogidos en las observaciones es la esperanza condicionada $E[y|X]$, tal que, $E[y|X]=X\beta$. Además, este supuesto de *regresores no estocásticos* permite demostrar que el vector de coeficientes MCO $\hat{\beta}$ es un estimador lineal insesgado de β.

Otro supuesto habitual del modelo de regresión lineal en relación con las perturbaciones aleatorias es aceptar que éstas están normalmente distribuidas, con media cero, como hemos visto, y varianza constante, σ^2. Aunque en este caso el supuesto de normalidad no es estrictamente necesario –puede asumirse una distribución de los errores alternativa–, resulta adecuado para la construcción de los contrastes estadísticos de validación de las estimaciones. Este supuesto de la varianza constante se complementa con el de independencia de las perturbaciones, generalmente conocido como supuesto de *no autocorrelación*, según el cual debe cumplirse que $\text{cov}[\varepsilon_i,\varepsilon_j|X]=0, \forall i\neq j$.

La propiedad de varianza constante se conoce como *homocedasticidad*, mientras que su vulneración recibe el nombre de *heterocedasticidad*. La razonabilidad de este supuesto queda condicionada a la tipicidad del análisis realizado. En la medida que en un conjunto de observaciones exista un sesgo de dimensión o de escala que afecte a los valores absolutos de los datos, la variabilidad de resultados recogida a través de la perturbación aleatoria también se verá afectada (p.e. la influencia del tamaño familiar en determinados gastos de consumo de los hogares).

Contraste de significación de los coeficientes estimados

El procedimiento más habitual para determinar en qué medida una variable exógena influye en la variable explicada, o lo que es lo mismo, hasta qué punto la participación de esa variable x_k resulta significativa (esencial) en terminos de probabilidad estadística en la formación de los valores de y , es la utilización del contraste estadístico basado en la t de Student, también conocido como t-ratio. Con su aplicación tratamos de contrastar si un parámetro determinado b_k que interviene en el modelo de regresión lineal estimado es significativamente diferente de cero. Este contraste de significación de los coeficientes de regresión es ofrecido por todos los programas informáticos econométricos.

La fundamentación teórica de este contraste está basada en el supuesto de independencia entre el vector de estimadores MCO $\hat{\beta}$ y el vector de

residuos . Como acabamos de ver, la condición de normalidad exigida a establece que su varianza será un valor constante s^2. Por tanto, para poder contrastar la hipótesis de significación necesitamos contar con una estimación del valor poblacional s^2. En la medida que b no es observable directamente, sino a través de la estimación , empleamos como estimador insesgado de la *varianza de la regresión*, calculada como,

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K} \qquad [34]$$

No olvidemos que σ^2 es un valor único, siendo s la desviación típica o *error estándar de la regresión*. Obtenido el valor s^2, podemos obtener los estimadores muestrales de la varianza muestral de cada uno de los parámetros estimados:

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \qquad [35]$$

Por tanto, el error estándar del estimador vendrá determinado por la raíz cuadrada del elemento *k*-ésimo de la diagonal principal de la matriz $(\mathbf{X'X})^{-1}$, que definimos como a_{kk}:

$$\hat{\sigma}_{\hat{\beta}_k} = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\Big|_{kk}} = \sqrt{\hat{\sigma}^2 \cdot a_{kk}} \qquad [36]$$

El contraste individual de significación más habitual consiste en comprobar si el parámetro b_k es significativamente distinto de cero, $H_1: b_k = 0$. Para ello, en primer lugar calculamos el estadístico *t* asociado al parámetro estimado *k* a partir de los datos proporcionados por la regresión:

$$t_{\hat{\beta}_k} = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \qquad [37]$$

Una vez obtenido el valor , se procede a comparar su valor absoluto, , con el valor teórico según tablas de la distribución *t* de Student, $t_{\lambda/2}$, valor que se define como el valor crítico porcentual de una distribución *t* de Student con $N - k$ grados de libertad.

Para comprender esta comparación hay que tener en cuenta que bajo el supuesto de distribución normal de e, el estadístico *t* correspondiente

al parámetro b_k sigue una distribución de tipo t de Student con $N - k$ grados de libertad, tal que,

$$t_{b_k} = \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 \cdot a_{kk}}}}{\sqrt{\frac{(N-k)\cdot\hat{\sigma}^2}{\sigma^2}}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 \cdot a_{kk}}} \qquad [38]$$

De este modo, para rechazar la hipótesis nula $H_0 : b_k = 0$, o lo que es lo mismo, que el coeficiente estimado $\hat{\beta}_k$ es *estadísticamente significativo*, debe cumplirse que el t-ratio definido en [37], en valor absoluto, sea mayor que el valor crítico en tablas $t_{\gamma 2}$. En ocasiones, la comparación de $|t_{b_k}|$ se hace con el valor 1,96, que es el que corresponde para una muestra grande y un grado de significatividad del 5%.

En la práctica, todos los programas que permiten la realización de estimaciones econométricas proporcionan directamente este contraste, sin necesidad de tener que acudir a las tablas estadísticas de la distribución t de Student. Normalmente, la verificación se realiza a través del cálculo de la probabilidad de aceptación de la hipótesis nula $H_0 : \beta_k = 0$, conocido generalmente como «P-valor». En nuestro caso, como vemos en la columna F, el módulo de Regresión Lineal de XLSTAT 2007 ofrece para cada estimador (variable) la probabilidad (Pr) de que el valor absoluto del t-ratio (columna E) –calculado como el cociente entre el valor del estimador (columna C) y la desviación típica correspondiente (columna D)– sea inferior al valor crítico $t_{\alpha/2}$, lo que supondría aceptar que $\beta_k = 0$, es decir, descartar la participación de dicha variable explicativa en el modelo.

	A	B	C	D	E	F	G	H
53								
54		Parámetros del modelo:						
55								
56		Fuente	Valor	Desviación típica	t	Pr > |t|	Límite inferior (95%)	Límite superior (95%)
57		Intersección	17531,631	2369,927	7,398	< 0,0001	12907,610	22555,652
58		EDAD	229,937	32,498	7,075	< 0,0001	161,045	298,829
59		GEN	-2241,292	826,756	-2,711	0,016	-3883,937	-498,647
60		EST	1927,036	378,949	5,085	0,000	1123,701	2730,372
61								
62								
63		Ecuación del modelo:						
64								
65		SAL = 17531,630794580						
66								

Habitualmente, para determinar la significación estadística de los coeficientes estimados suelen diferenciarse tres niveles de aceptación, en función de la probabilidad de rechazo del valor del coeficiente estimado que ofrece el programa. Así, se considera que cuando el valor de esta probabilidad (Pr) o P-valor es inferior al 1%, el coeficiente estimado es estadísticamente significativo al nivel del 1%. Cuando este P-valor (Pr) está comprendido entre 0,01 y 0,05, el coeficiente estimado es estadísticamente significativo al nivel del 5%, mientras que si el P-valor es superior a 0,05 pero inferior a 0,10, será estadísticamente significativo al nivel del 10%. Alternativamente, estas probabilidades pueden interpretarse en términos de la hipótesis no nula, $H_1 : b_k = 0$, de modo que en el primer caso, existe una probabilidad del 99% de que el coeficiente estimado participe en el modelo, del 95% en el segundo y del 90% en el tercero. Por debajo de este nivel de significación estadística aceptar la presencia del coeficiente resulta bastante cuestionable.

En la presentación de resultados de las estimaciones econométricas, la significación estadística al 1% suele marcarse con *** junto al valor del coeficiente, la correspondiente al 5% con **, y la significación estadística al 10% con *. Igualmente, se suele acompañar entre paréntesis el valor del estadístico *t* correspondiente al coeficiente estimado, lo que nos permite deducir la desviación típica a través de la expresión [37].

Para comprobar el funcionamiento de este contraste proponemos el siguiente ejercicio de regresión lineal, a realizar con las series de datos contenidas en la tabla 7. A partir de las mismas se trata de explicar el comportamiento ahorrador de los individuos en un año en relación con el momento de su jubilación («ahorro previsional»), considerando tres posibles variables explicativas –la edad del individuo, su renta anual y su expectativa respecto a su fortuna, recogida a través de su gasto en juegos de azar–.

	A	B	C	D	E	F
1						
2		Y(a1)	Y(a2)	X1	X2	X3
3		Ahorro Previsional	Ahorro Previsional	Edad	Renta Anual	Gasto Juegos Azar
4		512,0	448,44	24	24000	230
5		666,0	586,69	32	32000	329
6		1501,5	1428,75	38	78000	615
7		342,3	277,87	25	14500	170
8		678,0	597,49	56	31000	260
9		539,0	457,11	60	23000	200
10		261,0	191,69	32	9500	75
11		611,3	531,84	63	27500	450
12		279,0	209,69	32	10500	260
13		614,0	536,58	48	28000	420
14		216,3	151,97	25	7500	120
15		376,0	309,07	31	16000	90
16		727,0	647,98	52	34000	410
17						

TABLA 7

Los datos de la variable «Ahorro Previsional» se han obtenido mediante un proceso generador a partir de las variables X1 y X2. En el primer caso, *Y*(*a1*) se ha obtenido de forma determinística, a partir de la ecuación: $Y(a1)=50+1,25\cdot X1+0,018\cdot X2$. En el caso de *Y*(*a2*), hemos introducido un término aleatorio en la anterior ecuación:

$$Y(a1)=50+1,25\cdot X1+0,018\cdot X2+\varepsilon$$

Empleando el módulo de Regresión Lineal de XLSTAT 2007, en primer lugar estimamos el modelo determinístico, $Y(a1)=\beta_0+\beta_1\cdot X1+\beta_2\cdot X2$, con el propósito de comprobar los P-valores proporcionados para los coeficientes correspondientes.

Parámetros del modelo

Fuente	Valor	Desviación típica	t	Pr > |t|	Límite inferior (95%)	Límite superior (95%)
Intersección	50,000	0,000	0,000	< 0,0001	0,000	0,000
Edad	1,250	0,000				
Renta Anual	0,018	0,000				

Ecuación del modelo

Ahorro Previsional = 50 + 1,25*Edad + 0,018*Renta Anual

Como vemos, la estimación nos proporciona unos coeficientes exactamente iguales que los parámetros considerados en la generación de la variable dependiente. Consecuentemente, el test de significación de la *t* de *Student* proporciona una probabilidad nula de que los valores de los tres parámetros estimados sean distintos de los obtenidos en el proceso de regresión.

A continuación, repetimos el ejercicio pero para el modelo estocástico que incluye el término aleatorio $Y(a1)=\beta_0+\beta_1\cdot X1+\beta_2\cdot X2+\varepsilon$.

Parámetros del modelo

Fuente	Valor	Desviación típica	t	Pr > |t|	Límite inferior (95%)	Límite superior (95%)
Intersección	-2,288	0,077	-29,588	0,007	-3,317	-1,260
Edad	0,749	0,017	44,049	< 0,0001	0,711	0,787
Renta Anual	0,010	0,000	1.891,273	< 0,0001	0,010	0,010

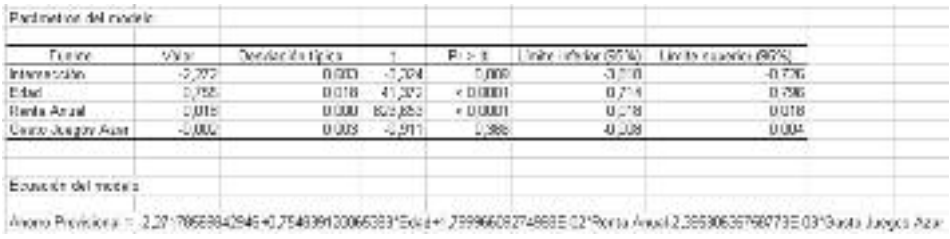
Ecuación del modelo

Ahorro Previsional = -2,288*0,749*Edad+0,010*Renta Anual

En los resultados que se alcanzan, se verifica esta significación estadística muy elevada para los tres parámetros, prácticamente del 99,99%, pues se obtiene un P-valor inferior al 1 por 1000 (Pr<0,0001) para las

dos variables explicativas, y un P-valor correspondiente al término independiente de 0,007. En este caso, el error de estimación es también muy bajo, pues la estimación del model proporciona un potencial explicativo a través del coeficiente de determinación R^2 del 99,99%. Hay que destacar que las diferencias en los valores de los coeficientes respecto del modelo determinístico atienden a la influencia de la perturbación aleatoria sobre la variable explicada, sin que esto deba interpretarse en términos de un posible error de predicción. Precisamente, estas diferencias en los valores estimados para los coeficientes b_0 y b_1 (el término independiente y la variable «edad»), que son los que más difieren de los valores determinísticos (-2,299 frente a 50, y 0,749 frente a 1,25), están capturando predictivamente el comportamiento de ese shock aleatorio. En cambio, vemos que el coeficiente b_2 correspondiente a la variable «renta anual» prácticamente toma el mismo valor (0,01798 frente a 0,018) que en el modelo determinístico.

Por último, procedemos a incluir en el modelo una tercera variable explicativa, la cantidad de dinero gastada anualmente en juegos de azar, para tratar de contrastar la conjetura que una mayor predisposición hacia este tipo de juegos reduce el comportamiento de ahorro previsional de los individuos: $Y(a1)=\beta_0+\beta_1\cdot X1+\beta_2\cdot X2+\beta_3\cdot X3+\varepsilon$. Puesto que se trata de un ejercicio con datos simulados en los que dicha variable no ha participado en la generación de los datos de la variable explicada, debemos esperar que el P-valor nos permita descartar el coeficiente correspondiente a dicha variable al no poder rechazar la hipótesis nula de que $\beta_3=0$.


Parámetros del modelo

Factor	Valor	Desviación típica	t	P-valor	Límite inferior (95%)	Límite superior (95%)
Intersección	-2,299	0,003	-1,204	0,009	-3,010	-0,776
Edad	0,750	0,018	41,203	< 0,0001	0,714	0,786
Renta Anual	0,018	0,000	823,853	< 0,0001	0,018	0,018
Gasto Juegos Azar	-0,002	0,003	-0,911	0,386	-0,008	0,004

Ecuación del modelo

Ahorro Previsional = -2,2978956842946+0,7548391200663831*Edad+1,7999660827486E-02*Renta Anual-2,2659063679873E-03*Gasto Juegos Azar

Los resultados de la estimación nos permiten observar cómo se cumple este pronóstico. El P-valor correspondiente a b_3 es de 0,386, lo que nos muestra un nivel de significación estadística muy bajo, de apenas un 60%. Hay que destacar que a pesar de que el número de observaciones es reducido –lo que influye en la determinación del valor crítico del estadístico *t*, el valor obtenido no llega a la unidad (0,911), cuando la significación en tablas al 90% requiere como mínimo un valor del estadístico de 1,282 (de 1,35 para una muestra de 13 observaciones).

Contraste de Significación de la Regresión

En el apartado anterior hemos visto cómo se puede contrastar si cada uno de los coeficientes estimados en el proceso de regresión es estadísticamente significativo. Se trata, sin duda, de uno de los contrastes esenciales para validar la estructura del modelo propuesto, pues determina la oportunidad de la participación de cada variable explicativa. No obstante, también podemos interrogarnos acerca de la significación estadística de la ecuación de regresión considerada de forma integral, es decir, contrastar si todos los coeficientes, con la excepción del término constante, son iguales a cero o no.

Es evidente que si cada uno de los coeficientes es nulo (las pendientes de la recta de regresión son cero), el coeficiente de correlación múltiple también deberá serlo. Por tanto, la forma natural de verificar esta hipótesis es a través del coeficiente de determinación R^2.

El método habitualmente empleado para contrastar esta hipótesis utiliza la distribución del estadístico conocido como *F* propuesto por Snedecor, definida como:

$$F(K-1,n-K)=\frac{\frac{R^2}{K-1}}{\frac{1-R^2}{n-K}} \qquad [39]$$

La idea de este contraste es la siguiente. Si aceptamos la hipótesis de que los coeficientes de regresión b_1, b_2, …, b_k (todos menos el correspondiente al término independiente, es decir, menos b_0), y las perturbaciones aleatorias recogidas en ε están normalmente distribuidas, el estadístico definido en [39] debe seguir una distribución *F* con $K - 1$ grados y $n - K$ grados de libertad. De acuerdo con la tabla de valores de la distribución *F*, valores muy elevados para $F(K - 1, n - K)$ nos llevan a rechazar la hipótesis propuesta.

En este sentido, si analizamos la expresión [39], observamos que para obtener valores elevados del estadístico el coeficiente R^2 debe mostrar también un valor elevado. Intuitivamente, este contraste global puede interpretarse como un indicador del error de ajuste que se cometería si admitiésemos que todos los coeficientes de las variables explicativas son nulos. Por tanto, si el valor del estadístico *F* definido de esta manera es elevado, podemos rechazar esa hipótesis, aceptando los valores de los coeficientes obtenidos en el proceso. Hay que aclarar que aunque un valor elevado del estadístico *F* es esperable que vaya acompañado de valores de la *t* de cada coeficiente que determinen su significación estadística, esto no tiene por qué producirse de forma inversa, pues puede demostrarse que ante coeficientes de cada variable

con elevada significación estadística, el conjunto de ellos puede no serlo. Un resultado de este tipo puede servir para identificar un escenario de *multicolinealidad* en las variables explicativas, lo que introduce un empeoramiento del proceso de ajuste de regresión en su conjunto.

Para la aplicación de este contraste, como vemos, el modulo de Regresión Lineal del programa XLSTAT 2007 ofrece directamente el valor de la expresión [39], y de nuevo, al igual que en el contraste de significación de los coeficientes, muestra el valor de la probabilidad de aceptación de la hipótesis nula, según la cual, $[b_1, b_2, \ldots, b_k]=0$.

Volviendo al ejemplo 2 –modelo de la ecuación de salarios con los datos de la tabla 4– observamos que el valor resultante para el estadístico *F* es de 48,929, lo que supone de acuerdo con los grados de libertad del modelo (16), un P-valor inferior al 0,0001 para aceptar la hipótesis de que todos los coeficientes de las variables explicativas fueran iguales a cero. Por tanto, contrastamos la significación estadística de todos los coeficientes considerados de forma conjunta al 99% (prácticamente el 100% puesto que la probabilidad es <0,0001).

Análisis de la varianza

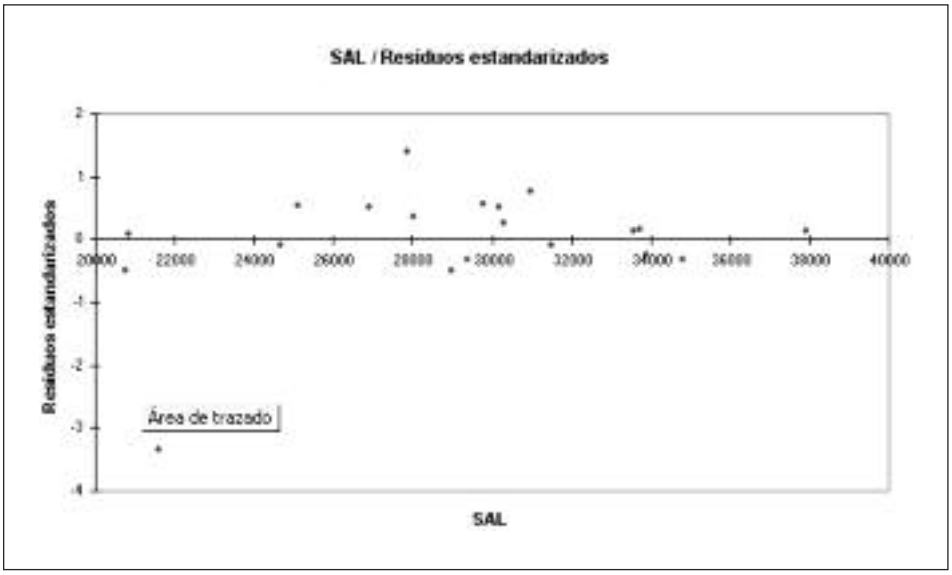
Fuente	GDL	Suma de los cuadrados	Media de los cuadrados	F	Pr > F
Modelo	3	382139948,929	127379982,976	48,929	< 0,0001
Error	16	41654125,209	2603382,826		
Total corregido	19	423794074,138			

Calculado contra el modelo Y=Media(Y)

La autocorrelación de las perturbaciones

Como hemos señalado en el capítulo, el carácter aleatorio de las perturbaciones resulta crucial a la hora de estimar un modelo de regresión. El carácter estocástico de las perturbaciones en el mundo real exige que los modelos estimados proporcionen un comportamiento estrictamente aleatorio de los residuos de la regresión. Este escenario deseable lleva a hablar de la existencia de *ruido blanco*, cuando analizamos gráficamente la evolución de los residuos de la regresión.

En este sentido, todos los programas econométricos, incluida la aplicación del programa que estamos utilizando, el XLSTAT 2007, ofrecen *plots* de residuos que permiten de forma aproximada apreciar si la distribución de los residuos ofrece alguna pauta funcional de comportamiento que nos aleje de ese supuesto ideal del *ruido blanco*. En la figura anterior observamos el *plot* correspondiente a los residuos (estandarizados) de la estimación de la ecuación de salarios del ejemplo 2.



Como vemos, a primera vista no encontramos un patrón de evolución que nos pueda hacer suponer la existencia de problemas de *heterocedasticidad*.

No obstante, la heterocedasticidad es un problema propio de las regresiones realizadas con datos de sección cruzada, especialmente en aquellas bases de datos dispuestas en forma de panel o pseudo-panel. Los patrones de comportamiento de las perturbaciones, en este caso, no tienen por qué presentar una explicación temporal. En cambio, en el tratamiento de las series temporales, este problema presenta una caracterización propia, conocida como la *autocorrelación* o existencia de *perturbaciones autocorrelacionadas*. La idea intuitiva es que la autocorrelación está generada por una relación en el tiempo tanto de las variables incluidas en el modelo como de aquellas otras omitidas. La relación temporal entre muchas de las variables económicas, especialmente las que integran las series macroeconómicas, inducen con frecuencia este tipo de fenómenos.

Aunque la econometría de series temporales constituye un área propia de desarrollo de la econometría, es habitual utilizar algún contraste estadístico para detectar fenómenos de autocorrelación. El contraste más habitual es el método conocido como contraste de *Durbin-Watson*. La noción básica de este contraste parte de la idea de que si las perturbaciones aleatorias de la información real están autocorrelacionadas, este hecho será identificado al detectar que existe autocorrelación de los residuos obtenidos en la estimación mínimo cuadrática.

El contraste de Durbin-Watson se define como:

$$DW = \frac{\sum_{t=1}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \qquad [40]$$

donde *e* representa el error de estimación en cada periodo temporal *t* (observaciones en series temporal).

Su aplicación es muy sencilla. Puesto que este estadístico se aproxima bastante al índice de autocorrelación muestral,

$$d = 2\cdot(1-\hat{\rho}) - \frac{e_1^2 + e_T^2}{\sum_{t=1}^{T} e_t^2} \qquad [41]$$

para muestras bastante grandes el segundo sumando de la expresión [41] será prácticamente nulo, por lo que el índice de autocorrelación será $d \cong 2\cdot(1-\hat{\rho})$. En consecuencia, si obtenemos en [40], con los residuos de la regresión, un valor de *DW* que difiera significativamente de 2, podemos rechazar la hipótesis nula de que no existe autocorrelación $(H_0 : \rho = 0)$.

Para aplicar este contraste teniendo en cuenta la significación estadística de la diferencia entre *DW* y el índice de autocorrelación *d*, el proce-

	Regresión de la variable SAL:	
	Coeficientes de ajuste	
	Observaciones	20,000
	Suma de los pesos	20,000
	GDL	16,000
	R^2	0,901711
	R^2 ajustado	0,883262
	MEC	2603382,825
	RMEC	1613,500179
	MAPE	3,390567
⟹	DW	2,099879
	Cp	4,000000
	AIC	298,993574
	SBC	302,966503
	PC	0,147433

dimiento consiste en determinar el valor crítico de *d* para la distribución según tablas. No obstante, se trata de contraste no exento de críticas, en la medida que la distribución de *d* depende de una cota inferior y otra superior determinada por los datos de la muestra, lo que en ocasiones da lugar a regiones de indeterminación en las tablas de Durbin-Watson.

La aplicación de Regresión Lineal del XLSTAT 2007, como cualquier otro programa de econométrico, ofrece el valor del estadístico *DW*. Como vemos, en el ejemplo 2 (estimación MCO de la ecuación de salarios), el valor del *DW* se aproxima a 2, por lo que podemos descartar con alta probabilidad la existencia de autocorrelación en las perturbaciones. No obstante, debemos recordar que la muestra empleada recoge información de corte transversal para un año determinado, lo que en cierto se opone a la naturaleza temporal del fenómeno.

BIBLIOGRAFÍA

Aznar A., García-Ferrer A. y A. Martín Arroyo (1994). *Ejercicios de Econometria I y Econometria II*. Madrid: Pirámide.

Greene, W. H. (1998). *Econometric Analysis*, 3.ª edición. Englewood Cliffs: Prentice Hall. Existe versión en castellano, *Análisis Econométrico*, 3.ª edición, Madrid, Prentice Hall, 1999.

Novales, A. (1993). *Econometría*, 2.ª edición. Madrid: McGraw-Hill.

Novales, A. (1996). *Estadística y Econometría*. Madrid: McGraw-Hill.

Wooldridge, J.M. (2006). *Introductory Econometrics*. A Modern Approach, 3.ª edición. Thompson. Existe traducción al castellano, *Introducción a la Econometría*. Un Enfoque Moderno, 2.ª edición. Madrid: Thomson-Paraninfo, 2006.

CAPÍTULO XVI
MODELO DE REGRESIÓN LOGÍSTICA BINARIA

CAROLINA NAVARRO RUIZ

16.1. INTRODUCCIÓN

Los modelos de regresión son modelos que habitualmente se utilizan porque nos permiten interpretar de forma simplifica la realidad sobre cualquier fenómeno que queramos estudiar. Por ejemplo, si quiero analizar y contrastar empíricamente qué factores son más relevantes en el crecimiento de un país, construimos un modelo de regresión donde incluiremos como variables explicativas aquellas que estimamos que van a influir sobre el crecimiento de un país y como variable dependiente aquella que resuma dicho crecimiento, por ejemplo, la renta per cápita. Un ejemplo más relacionado con la familia consistiría en analizar las causas que justifican el número de divorcios en España. Para ello utilizaríamos un modelo de regresión donde las variables explicativas serían aquellas que estimamos que explican las separaciones o divorcios de las parejas, como el número de años de duración del matrimonio, la edad de los cónyuges, el nivel económico, etc.

En el capítulo anterior hemos estudiado el modelo de regresión lineal donde la variable dependiente, como hemos visto en los ejemplos anteriores, es una variable continua. Sin embargo, en muchos casos el fenómeno que queremos estudiar no es continuo sino discreto, donde el objeto de nuestro interés consiste en un proceso de elección del individuo. Por ejemplo, podemos estar interesados en analizar la salud del sustentador principal de una hogar o el cabeza de familia y las características del individuo y su entorno que influyen en la misma. La buena o mala salud de un individuo vendrá explicada por factores tales como la edad, el sexo, los hábitos, su situación económica, el nivel de educación, etc. Este tipo de análisis exigen otro tipo de modelos distintos de los modelos de regresión lineal. Estos modelos son los que se denominan modelos de elección discreta, donde la variable dependiente consiste en una elección. En este capítulo vamos a estudiar los modelos de elección binaria, donde la variable dependiente es una variable dicotómica que toma dos valores, 0 y 1. Un ejemplo consistiría en analizar si una familia es pobre o no. La variable dependiente es una variable dicotómica que toma valor igual a 1 cuando el hogar se considera pobre y cero en caso contrario. Las variables explicativas que incluiríamos serían, el número de miembros del hogar, la composición del

hogar (si es monoparental, pareja con hijos o sin hijos, una persona sola, etc), el nivel de renta, la educación del cabeza de familia, su salud, si participa o no en el mercado laboral, etc.

16.2. EL MODELO

El modelo de elección binaria es una extensión de los modelos de regresión lineal múltiple, pero aplicado al caso de una variable dependiente dicotómica. Por este motivo, se exponen, en primer lugar, los argumentos por los cuales el modelo de regresión lineal no resulta apropiado cuando las variables son dicotómicas, para desarrollar, posteriormente, el modelo de elección binaria.

Brevemente, el modelo de regresión lineal se expresa como:

$$Y = X\beta + U \qquad [1]$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i \qquad [2]$$

donde u_i es el término de error. Uno de los supuestos es que el error, u, no está correlacionado con el conjunto de variables explicativas: $E(u|X_1, X_2, \ldots X_k) = 0$.

El lado derecho de la ecuación (2) puede tomar cualquier valor real debido a que no se ha impuesto restricción alguna sobre los parámetros de la estimación. También y_i puede tomar cualquier valor real. Cuando la variable observada y_i es dicotómica este modelo es inapropiado. En el caso de variables binarias, la atención debe centrarse en modelizar la probabilidad de que un individuo elegido aleatoriamente responda positivamente a la variable y_i. Esa probabilidad puede expresarse como una función lineal de las variables explicativas:

$$P_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i \qquad [3]$$

donde $P_i = P(y_i = 1|x)$.

La hipótesis de linealidad, sin embargo, está sujeta a dos límites importantes. Por una parte, el lado izquierdo de (3) es una probabilidad que toma valores comprendidos entre cero y uno, mientras que en el lado derecho no se ha impuesto ninguna restricción, por lo que puede tomar cual-

quier valor real. Por otra, cabe esperar que la tasa de cambio en la probabilidad de una respuesta positiva no será la misma para el rango total de *x*. En ese caso, una relación «curvilineal» puede ser más apropiada.

Para tener en cuenta dichas limitaciones, se necesita introducir un nexo de unión entre la probabilidad y las variables explicativas. Este nexo de unión debe proyectar el rango [0,1] en el rango (-∞,+∞) y debe tener forma de *s*. Los dos nexos de unión comúnmente utilizados en la práctica son las funciones logit y probit.

16.2.1. Logit y Probit

Los modelos Logit y Probit son los que habitualmente se utilizan como modelos de respuesta binaria. La función logística viene dada por:

$$F(\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki})=\frac{e^{\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki}}}{1+e^{\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki}}}$$

Supongamos que estamos interesados en analizar si una familia es pobre o no. Ello dependerá de un conjunto de variables, como por ejemplo, la composición el hogar, el sexo del sustentador principal (que es el individuo que aporta mayores ingresos al hogar), su nivel educativo, su situación en el mercado laboral, su salud, si mantiene o no relaciones sociales con las personas de su entorno, etc. La variable dependiente Y toma valor 1 si la familia es pobre y valor 0 en caso contrario.

La probabilidad de que una familia sea pobre o no, dadas las características del hogar y de su sustentador principal, es:

$$P_i=E[Y_i]=F(\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki})=\frac{e^{\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki}}}{1+e^{\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki}}}$$

La elección binomial entre las dos alternativas «ser una familia pobre o no» se especifica entonces a través de un modelo de regresión logístico binario:

$$Y_i^*=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki}+U_i^* \quad i=1,\ldots,n$$

donde Y_i^* es una variable no observable, denominada variable latente, que representa el estado de pobreza de la familia *i*, b el vector de parámetros correspondiente al vector *X* de variables explicativas y U_i^* el término de error.

Un factor de complejidad radica en que la variable Y_i^* no es observable. Tan solo se conoce que $Y_i = 1$ si $Y_i^* > 0$, $Y_i = 0$ si $Y_i^* \leq 0$. Para salvar esta restricción adoptamos el supuesto de que la función de distribución del término de perturbación aleatoria sigue una distribución logística:

$$P(Y_i = 1) = P(Y_i^* > 0) = P(X_i\beta > -U_i^*)$$
$$P(Y_i = 1) = P(X_i\beta \leq U_i^*) = F(X_i\beta)$$

donde F es la función logística. De este modo obtenemos la expresión del modelo logístico:

$$P_i = P(Y_i = 1) = \frac{e^{(X_i\beta)}}{1 + e^{(X_i\beta)}}$$

donde $X_i\beta$ indica la combinación lineal $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$

El modelo probit se especifica de manera similar al modelo logit pero utilizando como función de distribución la Normal:

$$P_i = E(Y_i) = F(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dt$$

16.2.2. Interpretación de los parámetros

En estos modelos, la estimación de los parámetros *b* no se puede interpretar directamente porque no representan necesariamente los efectos marginales del vector de regresores *X* sobre la variable dependiente. Al contrario de lo que ocurría en el capítulo anterior de regresión lineal múltiple, donde la estimación del parámetro *b* indicaba la estimación del cambio esperado en la probabilidad, por ejemplo, de ser pobre cuando la variable explicativa X_j cambia una unidad, en el modelo de elección binaria depende del valor inicial de la variable explicativa considerada.

La interpretación de los parámetros se puede realizar calculando la derivada de P_i respecto X_j.

La forma de calcular en el modelo logit los efectos marginales de cada variable explicativa debe recoger las particularidades de la distribución logística:

$$P_i = P(Y_i = 1) = \frac{e^{(X\beta)}}{1 + e^{(X\beta)}}.$$

Tomando logaritmos se obtiene que:

$$Ln(p/1-p) = X_i\beta$$

donde es fácil deducir que:

$$(p/1-p) = e^{X\beta}$$

La expresión $(p/1-p)$ representa los odds ratios, que es la relación entre la probabilidad de que un evento ocurra frente a la probabilidad de que no ocurra. Si la probabilidad de que ocurra el evento es mayor que la probabilidad de que no ocurra, entonces los odds ratios serán mayores que la unidad y el logit será mayor que cero, lo que indica una relación positiva, es decir, un aumento (disminución) en la variable independiente produce un aumento (disminución) en la variable dependiente. Cuando la probabilidad de que ocurra el evento es menor que la probabilidad de que no ocurra, entonces los odds ratios serán menores que la unidad y el logit será menor que cero, lo que indica una relación negativa, es decir, un aumento (disminución) en la variable independiente produce una disminución (aumento) en la variable dependiente. Si la probabilidad de que ocurra el evento es igual que la probabilidad de que no ocurra, entonces los odds ratios son igual a la unidad lo que indica que cualquier cambio en la variable independiente no tendrá ningún efecto sobre la variable dependiente.

Siguiendo nuestro ejemplo, supongamos que hemos estimado el modelo de regresión logística para calcular la probabilidad que tiene una familia de ser pobre en función de un conjunto de variables y el valor del coeficiente estimado de la variable independiente EDAD es -0,038. Entonces, $(p/1-p) = e^{-0,038} = 0,96,$ lo que significa que un aumento en una unidad (un año) en la edad del sustentador principal disminuye los odds de que una familia sea pobre en un 4% (100-96). Imaginemos ahora que el coeficiente estimado de la variable independiente GÉNERO, que toma valor 1 si es hombre y cero si es mujer, es 0,440, entonces $(p/1-p) = e^{0,440} = 1,55,$ lo que significa que los odds de que una familia sea pobre es 1,55 veces (o un 55%) mayor para las mujeres que para los hombres.

En el caso del modelo probit sólo señalar la gran complejidad en tratar de interpretar directamente los parámetros estimados, por lo que habitualmente se interpretan los signos que obtienen dichos parámetros. En ambos modelos, probit y logit, si el coeficiente estimado de un parámetro es positivo un aumento en el valor de la variable explicativa correspondiente produce un aumento en la probabilidad de escoger la opción igual a 1, en nuestro ejemplo, ser pobre. Es decir, que si por ejemplo el valor estimado del parámetro de la variable explicativa edad es positivo, significa que un aumento de la edad del individuo o sustentador principal provocará un aumento en la probabilidad de que ese hogar al que pertenece sea pobre. Si por el contrario, el valor del parámetro estimado de la variable explicativa correspondiente es negativo entonces, un aumento en el valor de la dicha variable explicativa provocará una disminución en la probabilidad de encontrarse en dicha opción. Es decir, en nuestro ejemplo, un signo negativo del coeficiente estimado de la variable explicativa de la edad significa que un aumento de la edad implica una reducción en la probabilidad de que la familia sea pobre.

La elección entre un modelo logit y un modelo probit depende del investigador. A efectos prácticos a menudo se elige el modelo logístico, sin embargo resulta muy complejo justificar la elección desde el punto de vista teórico. En Greene (1999) se señala que las dos distribuciones son similares, por tanto, tienden a dar probabilidades muy similares a valores intermedios de $B'X$. En cambio, este autor apunta que los resultados serán diferentes si la muestra que utilizamos como base de datos contiene pocas observaciones donde las variables dependientes sean igual a 1 o pocas igual a cero y exista gran variación en una de las variables que explican la variable dependiente. En general, los resultados que se obtiene de uno u otro modelo son muy similares. Como a efectos prácticos, la función logística es la que permite cálculos más sencillos es la que vamos a utilizar.

16.3. APLICACIONES PRÁCTICAS

Los datos que vamos utilizar para la aplicación empírica se encuentran en el archivo *echp.xls*. Es una muestra de 549 observaciones obtenidas a partir del European Community Household Panel (ECHP) correspondientes a la quinta ola (1998) en España. La aplicación práctica consiste en analizar la probabilidad de que un hogar sea pobre en función de un conjunto de características socioeconómicas. La variable dependiente es una variable cualitativa que toma valor 1 cuando el hogar es pobre y valor 0 en caso contrario. El resto de variables describen distintas características socioeconómicas del sustentador principal, edad, género, estado civil, nivel de educación, salud, si sufre enfermedades crónicas y la frecuencia con la que ve a parientes y amigos y características del hogar al que pertenece,

su composición y principal fuente de ingresos. Este conjunto de variables, incluidas en esta base de datos, se describen en el cuadro 1.

CUADRO 1. *Descripción de las variables del fichero* echp.xls

Variables	*Descripción*	*Categorías*
edad	edad del sustentador principal	—
género	género del sustentador principal	Varón (=1) Mujer (=2)
ecivil	estado civil del sustentador principal	Casado (=1) Separado (=2) Divorciado (=3) Viudo (=4) Soltero (=5)
edu	nivel de educación del sustentador principal	Tercer nivel, Universidad (=1) Segundo nivel enseñanza secundaria (=2) Sin estudios o menos de secundaria (=3)
salud	salud del sustentador principal (autoevaluación)	Muy buena (=1) Buena (=2) Regular (=3) Mala (=4) Muy mala (=5)
crónica	Si el sustentador principal sufre enfermedades crónicas	Sí (=1) No (=0)
social	Frecuencia con la que el sustentador principal se relaciona con otras personas	Mayoría de días (=1) Una o dos veces a la semana (=2) Una o dos veces al mes (=3) Menor frecuencia (=4) Nunca (=5)
fingresos	Prinicipal fuente de ingresos del hogar	Salario (=1) Empresario (=2) Pensiones (=3) Prestaciones por desempleo (=4) Otras prestaciones sociales (=5) Rentas de capital (=6)
chogar	Composición del hogar	Unipersonal (=1) Monoparental (=2) Parejas sin hijos (=3) Parejas con hijos (=4) Otro tipo de hogares (=5)
pobre	Ser pobre (estar por debajo del 60% de la mediana de la renta equivalente)	Sí (=1) No (=0)

La naturaleza de la variable dependiente y el tipo de análisis que se desea realizar exige la estimación de un modelo de regresión logística. XLSTAT dispone de una función para realizar regresiones logísticas.

El primer paso consiste en abrir el fichero que contiene la base de datos a partir del botón de la barra de herramienta *Archivo* → *abrir*. Seleccionamos el fichero de *echp.xls*.

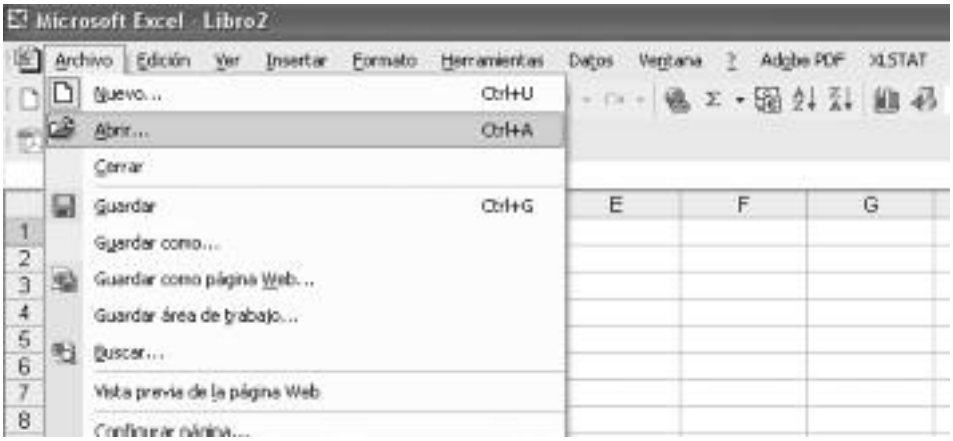


FIGURA 16-1

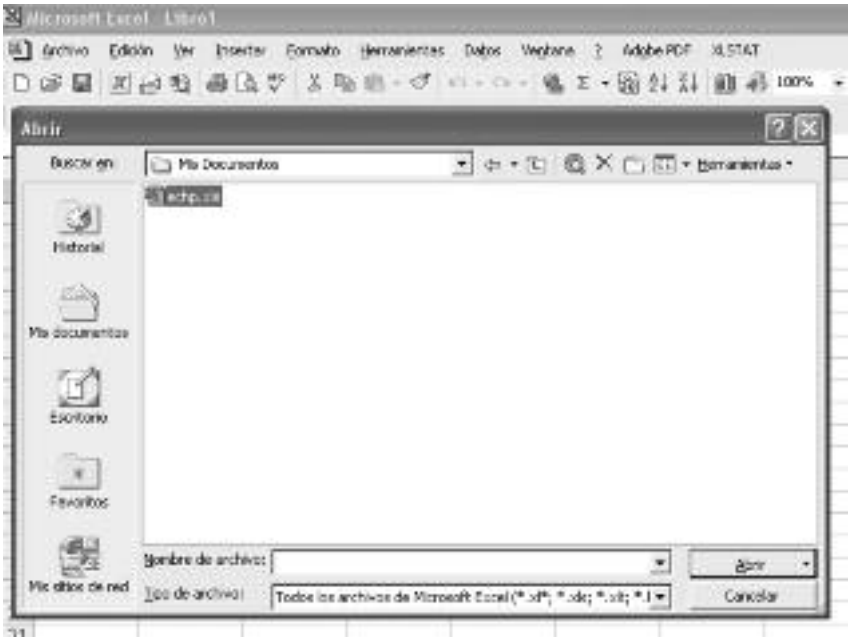


FIGURA 16-2

Una vez que hemos abierto el fichero de datos, en la barra de herramientas de Excel, hacemos clic en el botón *XLSTAT* y después seleccionamos la opción de *Modelación de datos* → *Regresión logística*. Si hacemos un clic en *Regresión logística* se abre un cuadro de diálogo con distintas pestañas correspondientes a cada una de las opciones de que dispone esta herramienta.

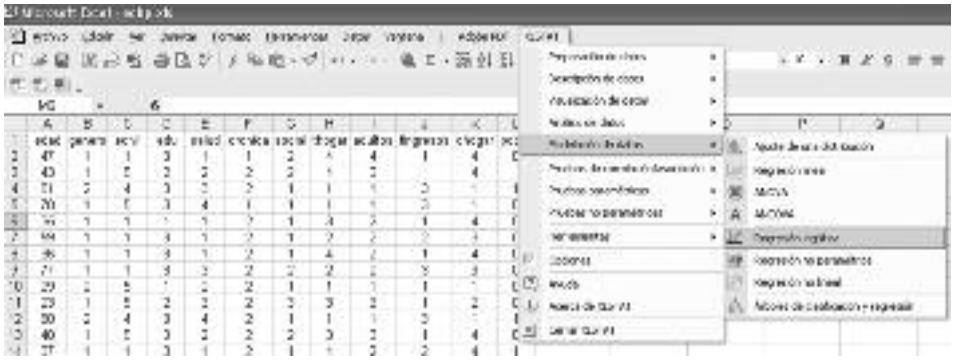
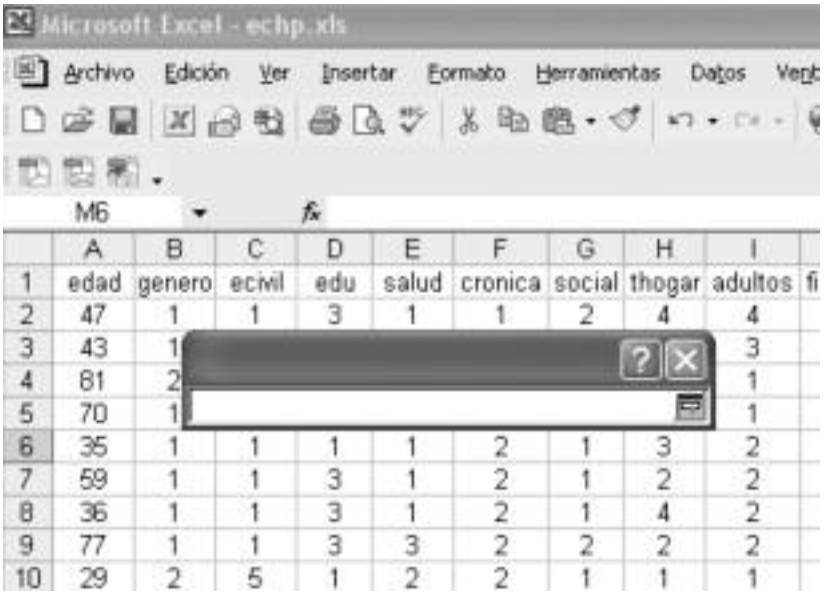


FIGURA 16-3

FIGURA 16-4

En el cuadro de diálogo (figura 16.4), en la pestaña *General* es donde realizamos la selección de la variable respuesta o variable dependiente, las variables cuantitativas así como las variables cualitativas. Previamente, para seleccionar variables debemos tener en cuenta en primer lugar, si en la primera fila de cada columna, donde están las variables, aparece el nombre de la variable. Si esto es así, como en el fichero de *echp.xls*, debemos seleccionar la opción *Etiquetas de las variables* en el cuadro de diálogo (figura 16.4) porque de esta forma XLSTAT identifica si la primera fila es un dato o una referencia, como en este caso. Para seleccionar la variable dependiente hacemos un clic en el desplegable de *Variable(s) respuesta* y automáticamente se abre una pequeña ventana y en el fondo tenemos nuestra página de excel con los datos. Para seleccionar la variable dependiente que en nuestra base de datos es la variable pobre, seleccionamos la columna de la variable pobre haciendo clic en la letra de la columna con el ratón. XLSTAT detecta automáticamente los límites de un cuadro de datos, localizando la presencia de una fila vacía. Cuando realizamos la selección observamos cómo en la ventana pequeña aparece los datos correspondientes a la selección realizada.



	A	B	C	D	E	F	G	H	I
1	edad	genero	ecivil	edu	salud	cronica	social	thogar	adultos
2	47	1	1	3	1	1	2	4	4
3	43	1							3
4	81	2							1
5	70	1							1
6	35	1	1	1	1	2	1	3	2
7	59	1	1	3	1	2	1	2	2
8	36	1	1	3	1	2	1	4	2
9	77	1	1	3	3	2	2	2	2
10	29	2	5	1	2	2	1	1	1

FIGURA 16-5

FIGURA 16-6

Para volver al cuadro de diálogo (figura 16-4) hacemos un clic en el botón derecho de la ventana pequeña. Para seleccionar las variables cuantitativas repetimos el mismo proceso que con la variable dependiente. Hacemos un clic en el desplegable de *Cuantitativas* y seleccionamos la columna de la variable edad, que es la variable cuantitativa que vamos a utilizar en nuestro ejemplo, haciendo clic en la letra de la columna con el ratón.

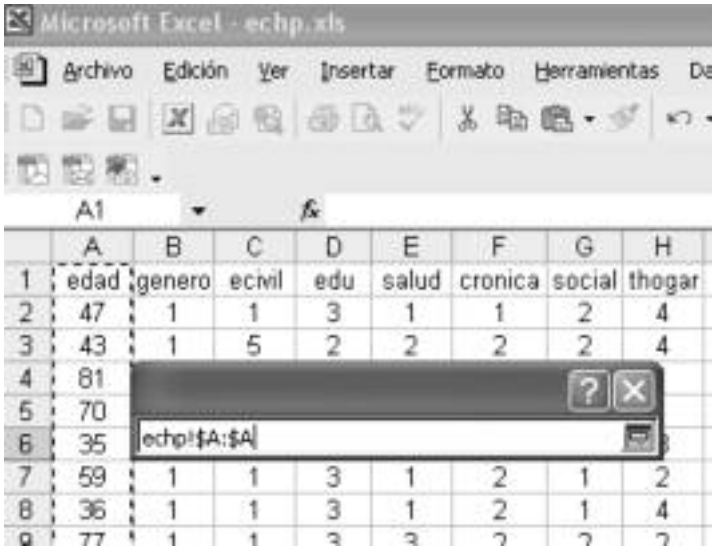


FIGURA 16-7

Para seleccionar las variables cualitativas, volvemos primero al cuadro de diálogo (figura 16.4) haciendo un clic en el botón derecho de la ventana pequeña. Las variables explicativas que vamos a incluir en la regresión logística son varias: las que representan el género, el estado civil, el nivel de educación y enfermedades crónicas del sustentador principal y la principal fuente de ingresos y la composición del hogar o familia a la que pertenecen. Se repite el mismo proceso que antes, pero en este caso, como son varias las variables a seleccionar, mantenemos presionada la tecla Ctrl durante la selección de las columnas correspondientes a estas variables (figura 16-8).

El campo de *Método* se refiere al método de regresión, que puede ser *Clásico*, que es el que realiza la regresión por Mínimos Cuadrados Ordinarios o *PCR*, que realiza la regresión por Componentes Principales. En el campo de *Modelo* existe un desplegable que nos ofrece las opciones de Logit, Probit, Log Log complementario y Gompertz. En nuestro ejemplo utilizamos el método clásico y el modelo logit y, por tanto, seleccionamos dichas opciones en los campos correspondientes del cuadro de diálogo.

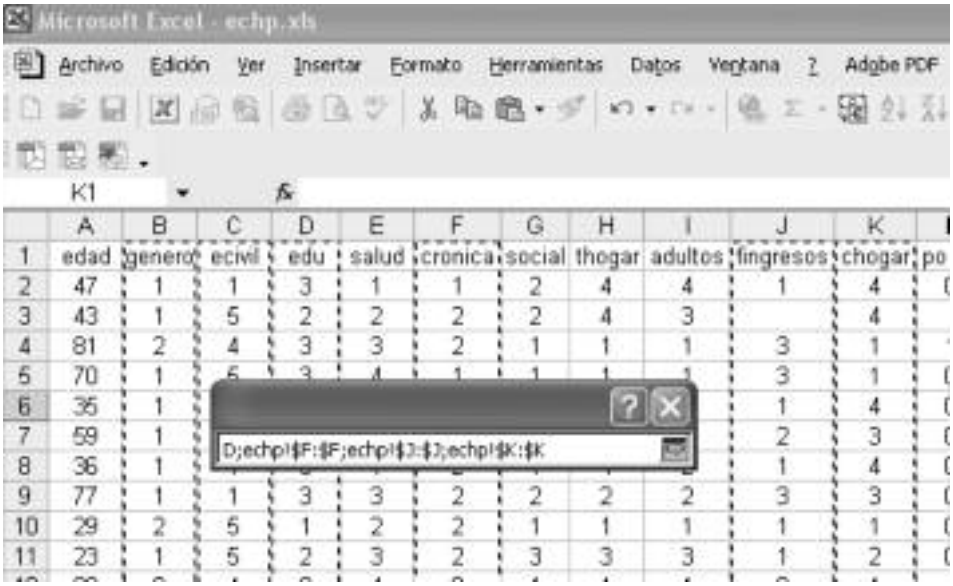


FIGURA 16-8

Las opciones correspondientes a los campos de *Rango*, *Hoja* y *Libro* se refieren al sitio donde se desea que aparezcan los resultados que se obtienen de la regresión logística. Si deseamos que los resultados obtenidos aparezcan a partir de una determinada celda de una hoja de Excel entonces seleccionamos la opción *Rango* en el cuadro de diálogo y después seleccionamos la cel-

da a partir de la cual queremos que figuren los resultados. Si queremos que los resultados aparezcan en una nueva hoja del libro de excel en el que estamos trabajando seleccionamos la opción *Hoja*. Por último utilizaremos la opción *Libro* si deseamos ver los resultados en un libro nuevo. En nuestro caso activamos la opción *Hoja* para que el programa nos muestre los resultados en una nueva hoja dentro del libro en el que estamos trabajando.

El campo de *Etiquetas de las observaciones* se utiliza cuando las observaciones de la base de datos disponen de etiquetas. En nuestro caso no disponemos de tales etiquetas por lo que XLSTAT genera automáticamente las etiquetas de las observaciones (Obs1, Obs2…).

El campo de *Pesos de las observaciones* sólo debe seleccionarse si el tipo de variable dependiente que se ha seleccionado, concretamente en el campo *Tipo de respuesta*, es la suma de variables binarias. En caso contrario, como es en nuestro ejemplo, en el que seleccionamos *tipo de respuesta binaria*, este campo está inactivo. El campo de *Pesos en la regresión* se utiliza cuando queremos ponderar las observaciones para ajustar el modelo. En nuestro caso no hemos incluido ponderaciones de las observaciones para realizar la regresión logística, por lo que no activamos esta opción.

Dentro del cuadro de diálogo, en la pestaña de *Opciones*, se puede determinar el nivel de Tolerancia (por defecto, el nivel por debajo del cual una variable es ignorada automáticamente es 0.001), el nivel al que se quiere realizar en su caso las interacciones y el intervalo de confianza para los diferentes test, entre otras opciones. En esta pestaña también se encuentra el campo de Selección del modelo. Si se desea utilizar un método de selección del modelo, este campo permite seleccionar entre: *Mejor modelo*, *Stepwise (Ascendente)*, *Stepwise (Descendente)*, *Ascendente* y *Descendente*.

FIGURA 16-9

El método de *Mejor Modelo* (figura 16.10) permite elegir el mejor modelo entre aquellos que se pueden obtener de la combinación de un mínimo de variables hasta un máximo de variables. También ofrece la opción de elegir entre distintos criterios para determinar el mejor modelo (Wald, etc.). El método *Ascendente* consiste en ir añadiendo variables empezando por la variable que mayor capacidad explicativa tiene. Es decir, en cada etapa se añade la mejor variable predictora aún no seleccionada. El método *Descendente* parte del conjunto completo de variables predictoras y va eliminando en cada etapa la peor variable predictora hasta que las variables que quedan en el modelo son todas pertinentes. Los métodos de *Stepwise (Ascendente)* y *Stepwise (Descendente)* son una combinación de las dos estrategias anteriores.

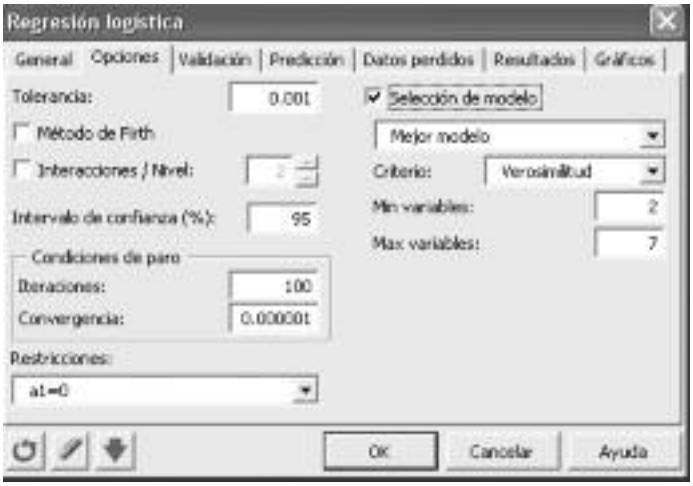


FIGURA 16-10

FIGURA 16-11

El *Stepwise (Ascendente)* comienza incluyendo la variable que más contribuye al modelo. Si la segunda variable es tal que su probabilidad de entrada está por encima del valor del umbral de entrada, se añade al modelo. Después de la tercera variable incluida se evalúa el impacto de eliminar cada variable a partir de la probabilidad umbral de retirada del modelo, de tal forma que si la probabilidad de retirada es mayor que el valor del umbral de retirada, esa variable se elimina del modelo. El método de *Stepwise (Descendente)* es similar al anterior pero partiendo de un modelo completo, con todas las variables.

Otra de las pestañas del cuadro de diálogo es la *Validación*, que permite usar una sub-muestra de datos para validar el modelo. Calcula predicciones para los datos de validación. En el campo de *Juego de validación* se ofrecen varias opciones para obtener las observaciones utilizadas para la validación: *aleatorio,* selecciona las observaciones aleatoriamente y debe especificarse el número de observaciones, *N últimas filas*, selecciona las *N* últimas observaciones y el número de observaciones debe especificarse, *N primeras filas*, selecciona las *N* primeras observaciones y debe especificarse su número o *variable de grupo*, que selecciona las observaciones que forman parte de un grupo que se selecciona a partir de una variable binaria que debe especificarse (figuras 16.12 y 16.13).

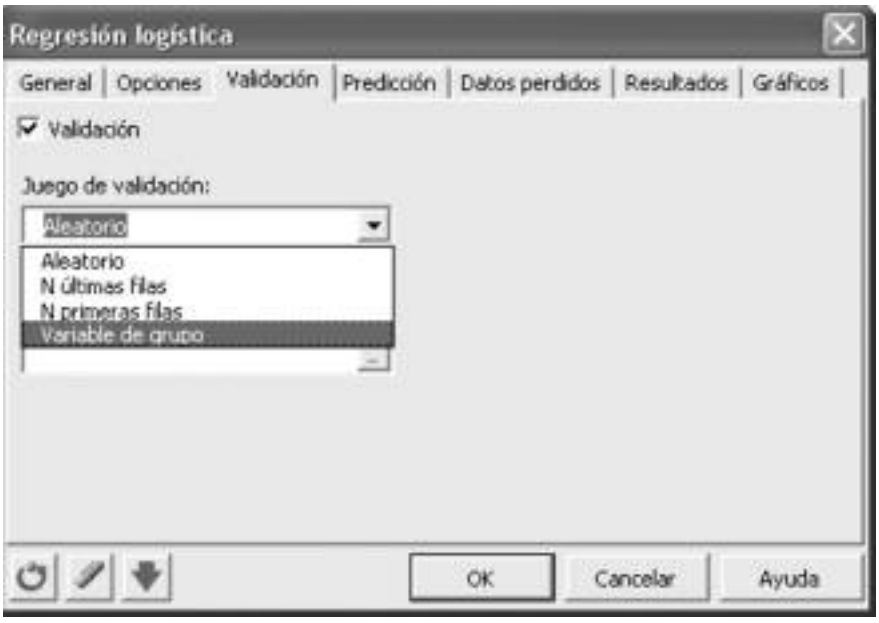


FIGURA 16-12

La pestaña *Predicción* del cuadro de diálogo permite calcular las predicciones para nuevas observaciones. Para utilizar esta opción es necesario asegurarse que los datos seleccionados están estructurados y ordenados de la misma forma que la realizada para la estimación. Es decir, hay que seguir el mismo orden que el utilizado en la selección de las variables para la estimación del modelo. Además, en la selección de las variables para la predicción no puede seleccionarse las etiquetas de las variables, si no que la primera fila de selección debe corresponder a los datos (figura 16.14).

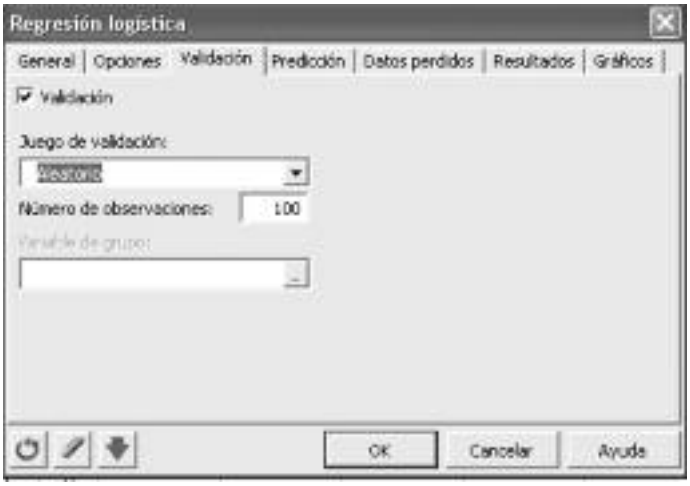


FIGURA 16-13

FIGURA 16-14

El cuadro de diálogo también incluye una pestaña denominada *Datos perdidos*. Una de las opciones que ofrece es eliminar las observaciones que contengan datos perdidos. Otra opción consiste en estimar los datos perdidos a partir de valor medio en el caso de variables cuantitativas y el valor modal en el caso de variables cualitativas u obtener el valor del vecino más próximo al valor perdido (figura 16.15).

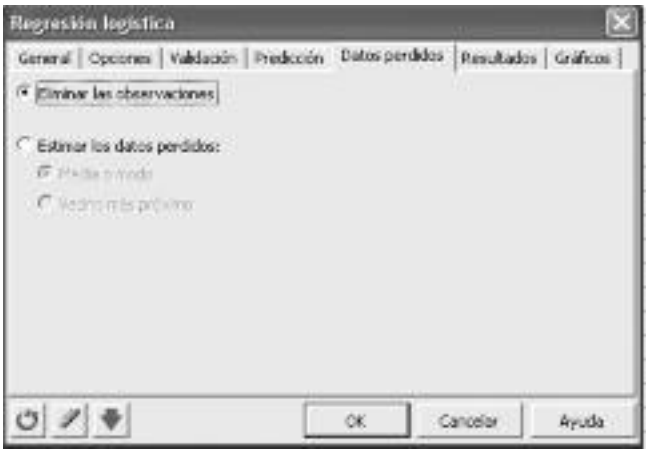


FIGURA 16-15

La pestaña de *Resultados* incluye distintas opciones: estadísticos descriptivos de las variables seleccionadas, matriz de correlaciones entre las variables explicativas, medidas de la bondad de ajuste del modelo, estimación de los parámetros del modelo, los coeficientes estandarizados, la ecuación del modelo, las predicciones y residuos de todas las observaciones, comparaciones, así como tablas de clasificación para evaluar el ajuste del modelo, entre otras.

FIGURA 16-16

Por último, este cuadro de diálogo incluye una pestaña de *Gráficos* para visualizar los gráficos de regresión (coeficientes estandarizados y predicciones) y la Curva ROC.

FIGURA 16-17

Cuando ejecutamos la regresión logística, XLSTAT muestra un amplio número de tablas y gráficos, según las opciones seleccionadas previamente, para analizar e interpretar los resultados. Los primeros resultados que observamos son los descriptivos. La primera tabla corresponde a las frecuencias de la variable dependiente, donde observamos que casi el 24% de los hogares son pobres.

Tabla 1

Variable	*Categorías*	*Frecuencias*	*%*
pobre	0	334	76,430
	1	103	23,570

Los descriptivos de la siguiente tabla corresponden a la variable explicativa cuantitativa incluida en el modelo.

Tabla 2

Variable	*Obs.*	*Obs. con datos perdidos*	*Obs. sin datos perdidos*	*Mínimo*	*Máximo*	*Media*	*Desviación típica*
edad	440	0	440	20,000	89,000	50,555	17,462

Observamos que esta variable no presenta ningún valor perdido, que la edad mínima de los individuos de los sustentadores principales de la base de datos es de 20 años, la edad máxima es de 89 años y la edad media es de 50 años.

Tabla 3

Variable	*Categorías*	*Frecuencias*	*%*
sexo	2	123	27,955
	1	317	72,045
ecivil	4	64	14,545
	5	81	18,409
	1	281	63,864
	3	5	1,136
	2	9	2,045
edu	3	298	67,727
	1	78	17,727
	2	64	14,545
cronica	2	328	74,545
	1	112	25,455
chogar	1	59	13,409
	4	184	41,818
	3	83	18,864
	2	48	10,909
	5	66	15,000
fingresos	3	118	26,818
	1	222	50,455
	2	53	12,045
	6	15	3,409
	5	20	4,545
	4	12	2,727

Los descriptivos de las variables explicativas cualitativas se presentan en una tabla de frecuencias. Observamos que el 72% de los sustentadores principales son varones y el 28% son mujeres. Respecto a su estado civil,

el 64% están casados, el 2% está separado, el 1% divorciados, el 15% son viudos y el 18% son solteros. En relación a su nivel educativo, observamos que el 18% han alcanzado el nivel educativo más alto (universidad), el 14% el segundo nivel de secundaria y el 68% han alcanzado un nivel menor de secundaria o no tienen estudios. Respecto a las enfermedades crónicas, una cuarta parte de los sustentadores principales sufre este tipo de enfermedades. Según la composición, el 13% de los hogares son unipersonales, el 11% monoparentales, el 19% parejas sin hijos, el 42% son parejas con hijos y el 15% son otro tipos de hogares (formados básicamente por adultos). La mitad de los hogares obtiene sus ingresos de los salarios, el 12% de actividades como empresarios, el 27% de las pensiones, el 3% de las prestaciones por desempleo, el 4% de otras prestaciones sociales y el 3% de rentas del capital.

En las opciones de regresión logística elegimos la opción de validación, que nos permitía usar una sub-muestra de datos para validar el modelo. Elegimos una sub-muestra de 100 observaciones de forma aleatoria. Las tablas de validación (tablas 4, 5 y 6) nos muestran que los descriptivos de la sub-muestra son muy similares a las presentadas en el modelo.

Tabla 4

Variable	*Categorías*	*Frecuencias*	*%*
pobre	0	72	73,469
	1	26	26,531

Tabla 5

Variable	*Obs.*	*Obs. con datos perdidos*	*Obs. sin datos perdidos*	*Mínimo*	*Máximo*	*Media*	*Desviación típica*
edad	100	0	100	18,000	89,000	50,760	18,302

Otra de las tablas que ofrece los resultados XLSTAT es la tabla de correspondencia entre las categorías de la variable respuesta (pobre) y las probabilidades (tabla 7). Observamos que a la categoría 0 (no pobre) le asigna la probabilidad 0 y a la categoría 1 (ser pobre) le asigna la probabilidad 1.

Tabla 6

Variable	*Categorías*	*Frecuencias*	*%*
sexo	2	26	26,000
	1	74	74,000
ecivil	4	18	18,000
	5	16	16,000
	1	65	65,000
	3	0	0,000
	2	1	1,000
edu	3	62	62,000
	1	25	25,000
	2	13	13,000
cronica	2	77	77,000
	1	23	23,000
chogar	1	18	18,000
	4	44	44,000
	3	22	22,000
	2	10	10,000
	5	6	6,000
fingresos	3	29	29,000
	1	54	54,000
	2	10	10,000
	6	3	3,000
	5	2	2,000
	4	2	2,000

Tabla 7

Categorías	*Probabilidades*
0	0
1	1

Las tablas 8 (coeficientes de ajuste), 9 (prueba de hipótesis nula) y 10 (análisis de tipo III) muestran las medidas de bondad de ajuste del modelo. La tabla 8 recoge el número de observaciones, la suma de las ponderaciones (en este caso como no hemos utilizado ponderaciones, por defecto el programa les asigna el valor 1, por lo que la suma de las ponderaciones coincide con la suma del número de observaciones), los grados de libertad y los estadísticos que se van a utilizar como medidas de la bondad de ajuste del modelo.

Estos estadísticos son los logaritmos de verosimilitudes asociados al modelo completo y al modelo independiente y los coeficientes R^2 (coeficientes que toman valores entre 0 y 1 y se interpretan de forma parecida al coeficiente de determinación R^2 de las regresiones lineales) de McFadden, Cox y Snell y Nagelkerke. Los coeficientes muestran el buen ajuste del modelo (los coeficientes R^2 son bajos pero el coeficiente de ajuste -2Log(Verosimilitud) muestra que el modelo se ajusta bien). Cuando decimos que los coeficientes muestran el buen ajuste del modelo estamos comparando los logaritmos de verosimilitudes del modelo completo y del independiente (modelo que no tiene ninguna variable explicativa, se reduce al término independiente). Para contrastar el buen ajuste del modelo estamos utilizando el siguiente contraste:

$$2\left[\ln L\left(\hat{\beta};X_{i,k}\right)-\ln L_0\left(\hat{\beta}^0;X_i,y_i\right)\right]$$

El estadístico que obtenemos, que sigue una distribución c^2 con 7 grados de libertad (la diferencia entre el número de parámetros del modelo independiente y el modelo completo) es superior al valor que corresponde en tablas, por lo que las variables explicativas (al menos una de ellas) son significativamente distintas de cero y por tanto, es mejor el modelo completo con todas las variables explicativas que el modelo independiente.

En la tabla 9 se contrasta que el modelo ajustado con todas las variables explicativas es significativamente más potente que el modelo independiente. El estadístico de Wald es mayor que el correspondiente en tablas con 18 grados de libertad y la probabilidad es menor de 0,05, por lo que se rechaza la hipótesis nula de que el modelo sea igual a la probabilidad que se obtiene mediante el modelo independiente (Y=0,230). Por tanto, se muestra el buen ajuste del modelo completo con todas las variables explicativas frente al modelo independiente.

Tabla 8

Estadística	*Independiente*	*Completo*
Observaciones	440	440
Suma de los pesos	440,000	440,000
GDL	439	421
-2Log(Verosimilitud)	474,079	329,139
R≈(McFadden)	0,000	0,306
R≈(Cox and Snell)	0,000	0,281
R≈(Nagelkerke)	0,000	0,000
AIC	478,079	367,139
SBC	486,253	444,787
Iteraciones	0	16

Tabla 9

Estadística	*GDL*	*Chi-cuadrado ajustado*	*Pr > Chi²≈*
-2 Log(Verosimilitud)	18	144,941	< 0,0001
Score	18	146,611	< 0,0001
Wald	18	91,729	0,000

Prueba de la hipótesis nula H0: Y=0,230 (Variable pobre).

Tabla 10

Fuente	*GDL*	*Chi-cuadrado ajustado (Wald)*	*Pr > Wald*	*Chi-cuadrado ajustado (LR)*	*Pr > LR*
edad	1	6,624	0,010	6,624	0,010
sexo-2	1	1,400	0,237	1,400	0,237
sexo-1	4	16,351	0,003	16,351	0,003
ecivil-4	2	5,238	0,073	5,238	0,073
ecivil-5	1	0,224	0,636	0,224	0,636
ecivil-1	4	15,694	0,003	15,694	0,003
ecivil-3	5	33,743	< 0,0001	33,743	< 0,0001

En la tabla 10 se contrasta el ajuste del modelo respecto a un modelo donde la variable de cada fila ha sido eliminada. La mayoría de las probabilidades correspondientes a cada una de las variables son menores de 0,05, por lo que la contribución de cada una de ellas al ajuste del modelo es significativa.

La tabla 11 presenta la estimación de los parámetros del modelo. Observamos que todas las variables explicativas incluidas en el modelo, salvo la variable de género y la de enfermedades crónicas, son significativas. La variable de estado civil muestra que los sustentadores principales que están casados o solteros tienen menor probabilidad de que su hogar sea pobre que aquellos hogares donde el sustentador principal es viudo. La variable edad muestra que la probabilidad de que un hogar sea pobre es menor cuanto mayor sea la edad del sustentador principal. El nivel de educación también contribuye a explicar la probabilidad de ser pobre. Los sustentadores principales cuyo nivel educativo alcanzado es el máximo (universidad) tienen menor probabilidad de que su hogar sea pobre que aquellos hogares cuyo sustentador principal no tiene estudios o tiene estudios inferiores a secundaria.

Tabla 11

Fuente	*Valor*	*Desviación típica*	*Chi-cuadrado de Wald*	*Pr > Chi²*	*Wald Límite inf. (95%)*	*Wald Límite sup. (95%)*
Intersección	4,373	1,338	10,678	0,001	1,750	6,996
edad	-0,038	0,015	6,624	0,010	-0,067	-0,009
genero-2	0,440	0,372	1,400	0,237	-0,289	1,168
genero-1	0,000	0,000				
ecivil-4	0,000	0,000				
ecivil-5	-2,241	0,557	16,168	< 0,0001	-3,333	-1,148
ecivil-1	-1,556	0,615	6,394	0,011	-2,762	-0,350
ecivil-3	-0,963	1,331	0,524	0,469	-3,571	1,645
ecivil-2	-17,937	1215,302	0,000	0,988	-2399,885	2364,011
edu-3	0,000	0,000				
edu-1	-1,057	0,520	4,139	0,042	-2,076	-0,039
edu-2	-0,690	0,511	1,820	0,177	-1,692	0,312
cronica-2	-0,168	0,355	0,224	0,636	-0,863	0,527
cronica-1	0,000	0,000				
chogar-1	0,000	0,000				
chogar-4	-0,363	0,680	0,284	0,594	-1,695	0,970
chogar-3	-1,596	0,725	4,846	0,028	-3,017	-0,175
chogar-2	-0,692	0,612	1,277	0,258	-1,891	0,508
chogar-5	-1,596	0,556	8,230	0,004	-2,686	-0,505
fingresos-3	0,000	0,000				
fingresos-1	-2,551	0,579	19,414	< 0,0001	-3,686	-1,416
fingresos-2	-1,868	0,625	8,949	0,003	-3,093	-0,644
fingresos-6	-0,602	0,720	0,700	0,403	-2,013	0,809
fingresos-5	-1,102	0,680	2,628	0,105	-2,435	0,230
fingresos-4	1,155	0,821	1,980	0,159	-0,454	2,765

También se observa que el tipo de familia influye en la probabilidad de ser pobres. Los hogares formados por parejas sin hijos o los que están formados por varios adultos tienen menor probabilidad de ser pobres que los hogares unipersonales. Por último, observamos que los hogares cuya principal fuente de ingresos proceden de los salarios o rentas de actividades empresariales tienen menor probabilidad de ser pobres que los hogares cuyos ingresos proceden básicamente de pensiones.

Tabla 12

Fuente	*Valor*	*Desviación típica*	*Chi-cuadrado de Wald*	*Pr > Chi²*	*Wald Límite inf. (95%)*	*Wald Límite sup. (95%)*
edad	-0,367	0,142	6,624	0,010	-0,646	-0,087
sexo-2	0,109	0,092	1,400	0,237	-0,071	0,289
sexo-1	0,000	0,000				
ecivil-4	0,000	0,000				
ecivil-5	-0,479	0,119	16,168	< 0,0001	-0,712	-0,245
ecivil-1	-0,412	0,163	6,394	0,011	-0,731	-0,093
ecivil-3	-0,056	0,078	0,524	0,469	-0,209	0,096
ecivil-2	-1,400	94,842	0,000	0,988	-187,287	184,488
edu-3	0,000	0,000				
edu-1	-0,223	0,109	4,139	0,042	-0,437	-0,008
edu-2	-0,134	0,099	1,820	0,177	-0,329	0,061
cronica-2	-0,040	0,085	0,224	0,636	-0,207	0,127
cronica-1	0,000	0,000				
chogar-1	0,000	0,000				
chogar-4	-0,099	0,185	0,284	0,594	-0,461	0,264
chogar-3	-0,344	0,156	4,846	0,028	-0,651	-0,038
chogar-2	-0,119	0,105	1,277	0,258	-0,325	0,087
chogar-5	-0,314	0,109	8,230	0,004	-0,529	-0,100
fingresos-3	0,000	0,000				
fingresos-1	-0,703	0,160	19,414	< 0,0001	-1,016	-0,390
fingresos-2	-0,335	0,112	8,949	0,003	-0,555	-0,116
fingresos-6	-0,060	0,072	0,700	0,403	-0,201	0,081
fingresos-5	-0,127	0,078	2,628	0,105	-0,280	0,026
fingresos-4	0,104	0,074	1,980	0,159	-0,041	0,248

La tabla 12 presenta los coeficientes estandarizados que se utilizan para comparar el peso relativo de las variables. Cuanto mayor es el valor absoluto de los coeficientes mayor es la importancia o el peso que tiene la variable. Esto también se puede analizar a través del gráfico de los coeficientes estandarizados (gráfico 1). Cuando el intervalo de confianza contiene el valor 0, el peso de la variable en el modelo no es significativo. Observamos que las categorías que antes hemos identificado como significativas se identifican también como tales a partir del intervalo de confianza alrededor de los parámetros estandarizados. Respecto al peso de cada una de ellas, observamos que la variable que mayor valor absoluto tiene es la segunda categoría de estado civil (separado). Esto se puede observar claramente en el gráfico 1. Sin embargo, esta categoría no es significativa, porque observamos que el intervalo de confianza alrededor de los coeficientes estandarizados contiene el 0.

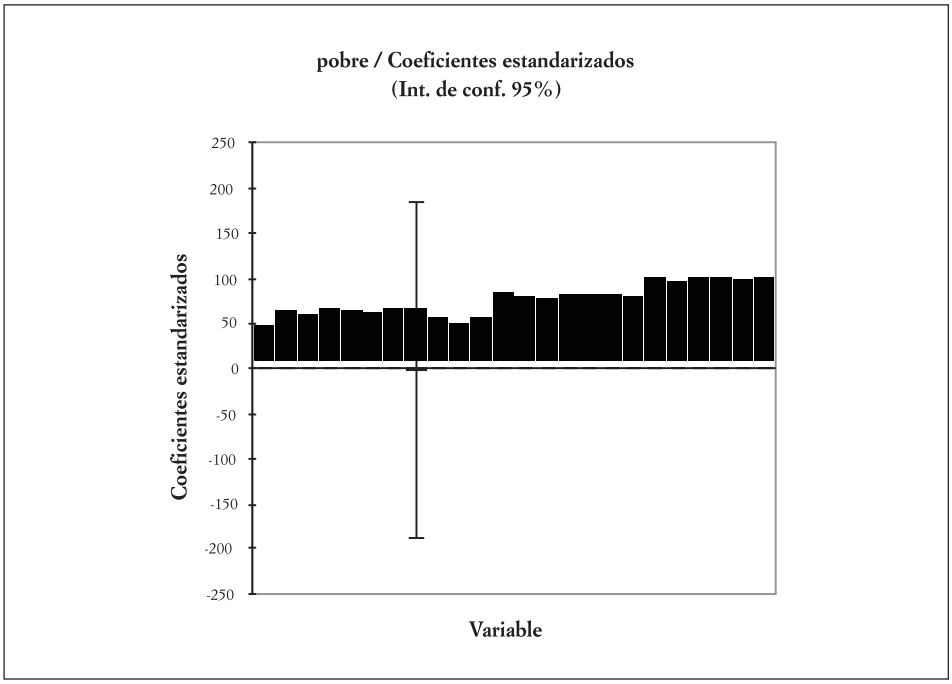


GRÁFICO 1

Los resultados también muestran la tabla de clasificación de la estimación para analizar el porcentaje de observaciones que han sido clasificadas en ambas categorías (pobre y no pobre) correctamente (tabla 13). Se comparan los valores observados para la variable dependiente (0 ó 1) con los valores predichos por el modelo. Al haber utilizado también la opción de validación, los resultados también muestran una tabla de clasificación para la sub-muestra de 100 observaciones (tabla 14). Las tablas de clasificación también muestran el buen ajuste del modelo: 85% y 82%, respectivamente.

Tabla 13

de \ a	*0*	*1*	*Total*	*% correcto*
0	324	15	339	95,58%
1	50	51	101	50,50%
Total	374	66	440	85,23%

La tabla 13 muestra que del total de hogares que no son pobres el modelo clasifica correctamente al 95% de las observaciones. Por otro lado, del to-

tal de hogares pobres clasifica como tales al 50% y al otro 50% lo clasifica como no pobres. En total clasifica correctamente al 85,23% de las observaciones.

La tabla 14, correspondiente a la opción de validación para una sub-muestra de 100 observaciones, muestra que el 82% de las observaciones son clasificadas correctamente. En particular, del total de hogares no pobres el modelo clasifica como tales al 94% y del total de hogares pobres clasifica como tales al 50% y al otro 50% como no pobres. Resultados muy similares a los anteriores.

Tabla 14

de \ a	*0*	*1*	*Total*	*% correcto*
0	68	4	72	94,44%
1	14	14	28	50,00%
Total	82	18	100	82,00%

La curva ROC muestra que el área por debajo de la curva es de 0,839, lo que significa que el modelo es bueno. Un valor de 1 sería un valor ideal, pero valores por encima de 0,7 indican que el modelo es bueno, valores entre 0,87 y 0,9 significa que es muy bueno y valores superiores a 0,9 muestran que el modelo es excelente.

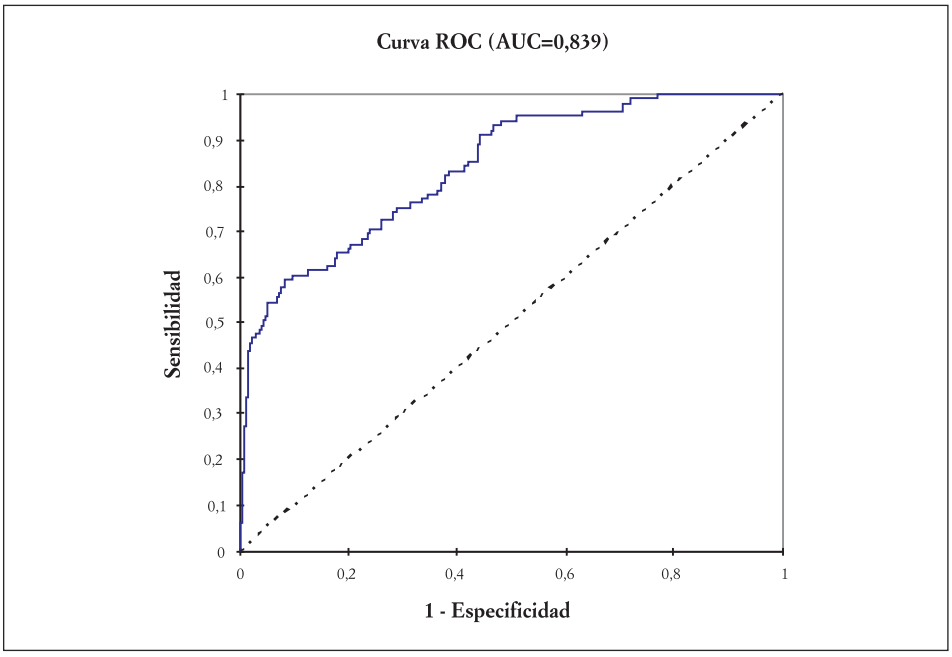


GRAFICO 2

Cuando el modelo incluye una o más variables explicativas, XLSTAT ofrece como resultado una tabla con los estadísticos correspondientes a los parámetros de las variables cualitativas tomadas en parejas (tabla 15).

Tabla 15

Contraste	*GDL*	*Chi-cuadrado ajustado*	*Pr > Chi²*
edad vs sexo-2	1	1,400	0,237
edad vs sexo-2	1	16,168	< 0,0001
edad vs sexo-1	1	6,394	0,011
edad vs ecivil-4	1	0,524	0,469
edad vs ecivil-5	1	0,000	0,988
sexo-2 vs sexo-1	1	1,476	0,224
sexo-2 vs ecivil-4	1	0,881	0,348
sexo-2 vs ecivil-5	1	0,000	0,990
sexo-1 vs ecivil-4	1	0,178	0,673
sexo-1 vs ecivil-5	1	0,000	0,989
ecivil-4 vs ecivil-5	1	0,000	0,989
edad vs sexo-2	1	4,139	0,042
edad vs sexo-1	1	1,820	0,177
sexo-2 vs sexo-1	1	0,300	0,584
edad vs sexo-2	1	0,224	0,636
edad vs sexo-2	1	0,284	0,594
edad vs sexo-1	1	4,846	0,028
edad vs ecivil-4	1	1,277	0,258
edad vs ecivil-5	1	8,230	0,004
sexo-2 vs sexo-1	1	6,510	0,011
sexo-2 vs ecivil-4	1	0,250	0,617
sexo-2 vs ecivil-5	1	4,785	0,029
sexo-1 vs ecivil-4	1	1,393	0,238
sexo-1 vs ecivil-5	1	0,000	1,000
ecivil-4 vs ecivil-5	1	2,185	0,139
edad vs sexo-2	1	19,414	< 0,0001
edad vs sexo-1	1	8,949	0,003
edad vs ecivil-4	1	0,700	0,403
edad vs ecivil-5	1	2,628	0,105
edad vs ecivil-1	1	1,980	0,159
sexo-2 vs sexo-1	1	2,149	0,143
sexo-2 vs ecivil-4	1	8,050	0,005
sexo-2 vs ecivil-5	1	4,092	0,043
sexo-2 vs ecivil-1	1	22,460	< 0,0001
sexo-1 vs ecivil-4	1	2,852	0,091
sexo-1 vs ecivil-5	1	1,004	0,316
sexo-1 vs ecivil-1	1	13,461	0,000
ecivil-4 vs ecivil-5	1	0,331	0,565
ecivil-4 vs ecivil-1	1	3,382	0,066
ecivil-5 vs ecivil-1	1	5,504	0,019

Los resultados que ofrece XLSTAT también muestran las predicciones y los residuos correspondientes a cada observación, pero por problemas de espacio (549 observaciones) no incluimos esta tabla.

BIBLIOGRAFÍA

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. y Tatham, R. L. (2006). Multivariate Data Analysis. Sixth Edition. Pearson, Prentice Hall. New Jersey.

Greene, W. H. (1999). Análisis econométrico. 3.ª edición. Prentice Hall, Madrid.

BREVE CURRICULUM DE LOS AUTORES

NURIA BADENES PLÁ

Es Catedrática de la Escuela de Empresariales desde 2002. Es Master en Hacienda Pública y Análisis Económico por el Instituto de Estudios Fiscales (1995) con Mención Especial. Doctorada en Ciencias Económicas por la Universidad Complutense de Madrid (2000), su tesis fue distinguida con el Premio Extraordinario y el Premio del Instituto de Estudios Fiscales. Ha realizado su labor investigadora colaborando con el Instituto de Estudios Fiscales, *L'Observatoire de l'Epargne Européenne,* la Fundación BBVA, FEDEA, FUNCAS, y como *Visitor Academic* de la *York University*. Obtuvo el Premio Círculo de Empresarios en 2002 junto al profesor González-Páramo. Sus áreas de interés incluyen redistribución, pobreza, microsimulación de reformas fiscales, educación, consumo de bienes nocivos, imposición marginal efectiva, activos financieros. Ha publicado sus trabajos en más de treinta artículos, libros y capítulos de libro.

M.ª TERESA LÓPEZ LÓPEZ

Licenciada en Ciencias Económicas y Empresariales por la Universidad Complutense de Madrid con la calificación de Sobresaliente, obteniendo en 1986 el grado de Doctor en Ciencias Económicas y Empresariales por la misma Universidad, con la máxima calificación. En 1987 realizó un Máster (Licenciatura Especial p.m.) en Economía Europea por el Instituto de Estudios Europeos, Universidad Libre de Bruselas. Ha sido Decana de la Facultad de Ciencias Económicas y Empresariales de la UCM desde mayo de 1998 hasta octubre de 2003. En la actualidad es miembro del Claustro de la UCM y de la Junta de Gobierno de la Facultad de CCEE de la misma Universidad. Pertenece al Departamento de Economía Aplicada VI de dicho centro desde 1980 y en la actualidad es Profesora Titular y dirige la Cátedra de Políticas de Familia UCM-AFA.

Las líneas de investigación en las que trabaja se centran en políticas públicas de carácter social con atención prioritaria en las relacionadas con la familia. Dirige el Grupo de Investigación «Políticas de Familia» en la Universidad Complutense, en el que participan once profesores de diversas disciplinas y pertenecientes a diferentes Universidades. Ha dirigido y participando en numerosas investigaciones relacionadas con políticas de familia, financiadas por instituciones públicas y privadas de ámbito nacional

e internacional. De las últimas realizadas destacan: *Políticas públicas de conciliación de la vida familiar y laboral en la Comunidad de Madrid. Un análisis de las variables del PHOGUE,* financiada por el Fondo Social Europeo en el marco de la Iniciativa Europea EQUAL 2002-2004; *Valoración de las políticas públicas destinadas a conciliar la vida laboral y familiar en la Unión Europea. Cuantificación de sus efectos en la economía social y laboral española,* incluida en el programa sectorial de estudios de I+D 99, finalizada en febrero 2003 y cuyos principales resultados se han publicado a finales del año 2004 por el Consejo Económico y Social.

Obtuvo el Premio de Investigación del Consejo Económico y Social 2000, que realizó en colaboración con investigadores de la Universidad de Sevilla y Pompeu Fabra.

Ha escrito y participado en más de 30 libros, entre los que destacan: *Familia y economía* (2007); *Políticas públicas de conciliación de la vida familiar y laboral en la Unión Europea* (2004). *El trabajo asalariado como generador de derechos: maternidad, enfermedad, jubilación y desempleo. Diferencias entre hombres y mujeres* (2003). *Políticas Públicas y Familia* (2005); *Protección social a la familia en los Estados de la Unión Europea* (1999). Igualmente ha escrito más de 50 artículos entre los que se encuentran: *Políticas públicas de familia: fundamentos y propuestas* (2004). *Políticas públicas de protección social a la familia en la Unión Europea* (2002). *La estructura actual de la familia* (2002). *Bases para la orientación de las políticas sociales. Referencias y propuestas en el ámbito fiscal* (2001).

Está casada y tiene tres hijos.

CAROLINA NAVARRO RUIZ

Es Doctora en Ciencias Económicas y Empresariales (Universidad Complutense de Madrid), Premio Extraordinario 2003/2004, Licenciada en Ciencias Económicas y Empresariales (Universidad Complutense de Madrid). Es profesora del Departamento de Economía Aplicada y Gestión Pública de la UNED, donde desarrolla su actividad docente e investigadora, especialmente en el área del análisis del bienestar social. Cuenta con una amplia experiencia docente desde 1998 hasta la actualidad en diferentes universidades y centros, en distintas licenciaturas, doctorado y cursos de postgrado oficial, impartiendo un amplio repertorio de asignaturas. Es investigadora externa del Instituto de Estudios Fiscales (Ministerio de Hacienda) desde 2001 y ha participado en diversos proyectos derivados de convocatorias públicas y competitivas, congresos nacionales e internacionales y publicaciones científicas, entre las que cabría destacar un libro publicado por el Consejo Económico y Social que recibió un Accesit en su convocatoria 2004, y dos artículos, uno publicado en Journal

of Housing Economics y otro en Applied Economics. Ha obtenido becas predoctorales en convocatorias públicas competitivas, nacionales y europeas. Ha realizado estancias de formación investigadora en distintos centros en el extranjeros (London School of Economics and Political Science y University of Essex), así como numerosos cursos de especialización en España y en el extranjero. Respecto a la experiencia profesional, cabe destacar la realización de numerosos informes técnicos y proyectos relacionados con la evaluación de políticas publicas para distintas instituciones públicas y privadas.

JORGE ONRUBIA FERNÁNDEZ

Nació en Madrid en 1963. Licenciado en Ciencias Económicas y Empresariales por la Universidad Complutense de Madrid y Doctor en Economía por esta misma Universidad. Profesor Titular del Departamento de Hacienda Pública y Sistema Fiscal de esta Universidad, en el que enseña Economía Pública desde 1988 y del que actualmente es Director. Ha sido investigador en plantilla del Instituto de Estudios Fiscales del Ministerio de Economía y Hacienda desde 1991 hasta 2001, y Vocal Asesor del Director General para Investigación desde 2001 hasta 2003. Ha participado en proyectos de investigación para la Fundación de las Cajas de Ahorros, para la Fundación BBVA, para la Fundación Modernización de España y ha actuado como consultor para el Banco Mundial y el Fondo Monetario Internacional en materia de Política Fiscal y Reglas Fiscales. Actualmente es miembro del Consejo de Redacción de *Hacienda Pública Española/Revista de Economía Pública*, publicación de la que ha sido Secretario del Consejo Editor entre 2001 y 2003, y Editor ejecutivo de *e-publica*. Sus principales líneas de investigación son la economía de la imposición, la distribución y redistribución de la renta, el análisis económico de la gestión pública, la economía de la vivienda y los aspectos institucionales de la política fiscal. En todas ellas tiene un número importante de publicaciones tanto de ámbito nacional como internacional.

CÉSAR PÉREZ LÓPEZ

Licenciado en Matemáticas (especialidad de Estadística) por la Universidad de Valladolid y Licenciado en CC. Económicas por la UNED. Pertenece al Cuerpo Superior de Estadísticos del Estado y al Cuerpo Superior de Sistemas y Tecnologías de la Información de la Administración del Estado. Actualmente es Vocal Asesor (Unidad de Estadística) en el Instituto de Estudios Fiscales y Profesor Asociado en el Departamento de Estadística e Investigación Operativa III de la Universidad Complutense impartiendo docencia en la Escuela de Estadística. Experiencia laboral en otras

Instituciones como el Instituto Nacional de Estadística y la Agencia de Protección de Datos. Es autor de numerosas publicaciones (libros y artículos) de referencia en los campos de la Estadística, Economía, Matemáticas e Informática.

DANIEL SANTÍN GONZÁLEZ

Daniel Santín González (Madrid, 1974) es Licenciado y Doctor en CC. Económicas y Empresariales por la Universidad Complutense de Madrid (UCM). Es además *Experto* en Análisis de Datos en Investigación Social por la E. U. de Estadística de la UCM. Sus áreas de interés son la economía de la educación, la eficiencia y la productividad en el sector público y las aplicaciones económicas de las técnicas estadísticas de análisis multivariante y *Minería de Datos*. Ha realizado diversas estancias de investigación en universidades extranjeras de prestigio, entre ellas, el *Department of Econonics* de *Harvard University (Estados Unidos)* y el *Centre de Recherche en Economie Publique et de la Population (CREPP)* de la *Université de Liège* (Bélgica). Ha publicado distintos libros y artículos científicos en revistas nacionales e internacionales de prestigio. Ha sido además Premio Extraordinario de Doctorado y Segundo Premio Nacional a la Investigación Educativa del Ministerio de Educación en la modalidad de Tesis Doctorales. Actualmente es Profesor Contratado Doctor en el Departamento de Economía Aplicada VI (Hacienda Pública y Sistema Fiscal) de la UCM.

AURELIA VALIÑO CASTRO

Doctora en Ciencias Económicas y Empresariales de la Universidad Complutense de Madrid, bajo la dirección del Profesor D. Enrique Fuentes Quintana, con la calificación de «cum laude». Es Profesor Titular del Departamento de Economía Aplicada en la Facultad de Ciencias Económicas y Empresariales de la Universidad Complutense de Madrid (U.C.M.), a la cual pertenece desde 1982. Ha sido Vicedecana de la Facultad de Ciencias Económicas y Empresariales de la UCM desde 1995 hasta 2003, ocupando el cargo de Decana durante los dos últimos meses de este periodo.

Su investigación se dirige hacia el campo de la evaluación de políticas públicas de la familia, con especial énfasis en educación, vivienda y conciliación de la vida familiar y laboral, así como economía de la defensa y gasto público en seguridad. Ha publicado 20 libros y capítulos de libros y más de 50 artículos en revistas nacionales e internacionales de las que destacan Applied Economics, Defence and Peace Economics, European Economy, Hacienda Pública Española, Papeles de Economía Española, Información Comercial Española, entre otras. ha colaborado en investigaciones con la

Secretaría de Estado de Hacienda, Ministerio de Trabajo, Ministerio de Educación, el Instituto de Estudios Fiscales, Fundación de la Confederación de Cajas de Ahorros (FUNCAS), la Comunidad de Madrid, La Fundación Europea Sociedad y Educación, Fundación Acción Familiar, la Fundación BBVA, la Comunidad Europea y otras instituciones relevantes. Recibió el premio de investigación de la Secretaría de Estado de Hacienda de 1990 y fue miembro del equipo de investigación que recibió el premio del Consejo Económico y Social del año 2000.